Part IV

User Experiment

In Part III I described two implicit feedback frameworks: one heuristic-based and one probabilistic. Both approaches used searcher interaction with document representations to generate new query statements and estimate changes in the information needs of searchers. The part concluded with a simulation-based evaluation of different candidate implicit feedback models, including parts of the heuristic-based and probabilistic frameworks from earlier chapters. The probabilistic model based partly on Jeffrey's rule of conditioning performed best and was therefore selected as part of the experiment now presented. The experiment tests the value of the framework in detecting current information needs and tracking them over a search session, and the effectiveness of different types of interface support to communicate its decisions. Unlike the tests carried out in Part III, this experiment involves human subjects, and in addition to testing the probabilistic framework this experiment evaluates how much control searchers really want over in their interaction with the implicit feedback framework though the provision of relevance information, query reformulation and making search decisions.

Chapter 9

Experimental Methodology

9.1 Introduction

The simulation-based study in the previous chapter tested how well implicit feedback models improved search effectiveness and 'learned' what information was relevant. The study found that the term selection model based on Jeffrey's rule of conditioning outperformed the other models tested in a variety of information seeking contexts. In this chapter the value of the probabilistic framework described in Chapter Seven (of which the Jeffrey's Conditioning Model is part) is tested with human subjects. The framework includes components to estimate information needs and track changes in them over a single search session. The experiment also evaluates different forms of interface support for presenting the decisions the framework makes. Three search interfaces are evaluated that vary the amount of control searchers have over creating queries, providing relevance indications and making search decisions. In this chapter I describe the methodology used to evaluate the probabilistic framework and interface support mechanisms in all experimental systems. The chapter begins by describing two pilot studies, and then further describes the experimental methodology.

9.2 Pilot Testing

Two pilot tests were carried out prior to this experiment: one tested the a prototype content-rich interface and the heuristic-based framework described in Chapter Six, the second debugged the questionnaires and search tasks used the experiment described in this chapter. In the remainder of this section I describe each of these tests.

9.2.1 Pilot Test 1: Interface and Heuristic-based Framework

The first pilot test evaluated a prototype system developed based on the content-driven principles described in Part II. This tested the interface support mechanisms and the effectiveness of the heuristic-based implicit feedback framework described in Chapter Six. Two experimental interfaces were created and 24 experimental subjects were recruited. This test allowed me to evaluate a prototype version of the interface used in the experiment described later in this chapter. As a result, I resolved interface design issues, obtained a better understanding of subject interaction with such interfaces, and established the effectiveness of the heuristic-based implicit feedback framework. This test is described in more detail in Appendix D.

9.2.2 Pilot Test 2: Questionnaires and Search Tasks

This second pilot study debugged the questionnaires and the search tasks used in this experiment. Minor changes to the wording of questions in the questionnaires were made as a result of subject feedback. However, the main aim of this pilot test was to investigate the suitability and complexity of the search topics. In the main experiment subjects are required to choose three search tasks, one of high complexity, one of moderate complexity and one of low complexity. Subjects were presented with three task sheets, each containing six tasks on six topics. Subjects chose a task from each sheet, but could not choose the same topic more than once.

Borlund (2000b) suggested the most important factor in a good simulated situation was the degree to which the topic engaged the subject's interest. Allowing subjects to choose tasks gave them more control over the search situation they were engaged in than simply allocating tasks to them on an arbitrary basis. In Pilot Test 1 I found that the level of interest in the search topic was the most important factor for experimental subjects when choosing one task over other alternatives.

Prior to starting the experiment, the task sheets were given to six randomly chosen volunteers. The volunteers were asked to read each of the tasks, place themselves in the simulated search scenario, and comment on the clarity and complexity of the task. These comments were informal and are not reported in this thesis. However, they did motivate slight changes in the wording of some tasks. In general, feedback on task complexity matched the categorisation used when developing the tasks. This was tested further in the main experiment and results are reported in later chapters.

In this section I have described two pilot tests that evaluate a prototype of the systems used in this experiment and debugged the questionnaires, search tasks and experimental procedures. In the remainder of this chapter I describe the methodology for the main experiment, beginning in the next section with the experimental systems.

9.3 Experimental Systems

Three experimental systems were developed to test these hypotheses. These systems varied in three ways: *relevance indication*, *query formulation* and *retrieval strategy selection* and used variations of interface components tested already in this thesis. A 'Checkbox' system (S_{Check}) allowed searchers to mark relevant items and use the items marked to create new queries. A 'Recommendation' system (S_{Recomm}) suggested additional query terms and retrieval strategies based on implicit relevance indications gathered from searcher interaction. An 'Automatic' system (S_{Auto}) automatically creates a new query and chooses the most appropriate retrieval strategies. No system gave subjects complete control over the terms used and search decisions taken. That is, all systems offered assistance in creating new queries, choosing how to use these queries, or both activities. Previous studies in IR have demonstrated that systems that offer feedback outperform systems where searchers are solely responsible for interaction decisions (Koenemann and Belkin, 1996; Beaulieu, 1997). I therefore felt it was unnecessary to include such a system did not offer any support in this experiment. These systems are described in more detail in Chapter Ten.

9.4 Equipment

I controlled the experiment from a laptop computer. The experimental systems ran on this computer and I sat next to computer for the duration of the experiment. An additional 21 inch monitor, a standard QWERTY keyboard and two-button optical mouse were connected to the laptop. [30] The experimental subject used these standard devices rather than those on the laptop, as shown in Figure 9.1. I felt these devices were more familiar to subjects than those on the laptop, which had a smaller display, a smaller keyboard and a touchpad for controlling the mouse pointer.

[30] The laptop computer had an AMD Athlon 2.4 GHz processor with 512 MB of RAM. The operating system was Microsoft Windows XP Professional and the Web browser used was Internet Explorer 6.0. All applications were written in Java, Dynamic HTML and JavaScript.

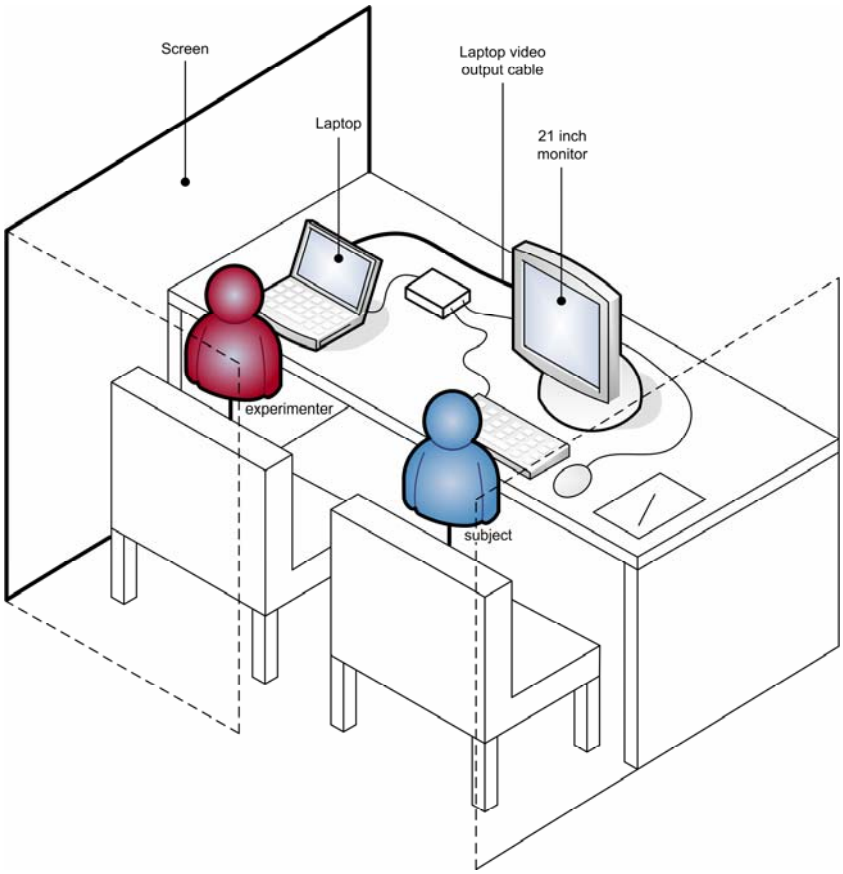


Figure 9.1. Equipment setup for the experiment.

Screens were positioned on three sides of the experimental location to block off noise and other distractions. I used the laptop to control the setup of experimental systems, control the construction of interaction log headers (described in Section 9.11) and observe subject interaction in an unobtrusive way. This also allowed me to intervene should there be any problems with the experimental systems. This intervention was limited only to occasions where technical problems prevented the subject from continuing with their search; I offered no other support.

9.5 Document Domain

The World Wide Web was used as the document domain for this experiment since subjects had experience interacting with Web documents, effective search systems were readily available and realistic search scenarios could be easily created. No restrictions were placed on the type of document that could be viewed or how far away from the experimental systems' result interface the subjects could browse. Restrictions were placed on whether external search systems (e.g., Google) could be used. These were seen as replacements for the experimental systems and were not permitted. Subjects were allowed to search within a

document using the 'Find' function of the Internet Explorer browser. Many subjects used this function to locate keywords within a Web document.

9.6 Subjects

The experimental subjects were mainly staff and undergraduate and postgraduate students at the University of Glasgow. 48 subjects were recruited. Half were male and half were female. Subjects were paid £12 (approximately €18) for participating. In this section I describe how volunteers were recruited and how the final set of subjects was selected.

9.6.1 Recruitment

Recruitment was targeted at two groups of subjects; *inexperienced* and *experienced*. In a related study, Holscher and Strube (2000) showed that experienced and novice Web searchers conduct their searches differently. Since the Web has a heterogeneous user population it is important to investigate how well the techniques I propose perform for different subject groups. I define the subject groups as:

i. **Inexperienced:** infrequent computer users, inexperienced searchers.
ii. **Experienced:** frequent/professional computer users, experienced searchers.

Subjects were not classified into their groups until after they had completed an 'Entry' questionnaire that asked them about their search experience and computer use. Subjects were recruited using electronic mails and advertisements per the ethics code of the Faculty of Information and Mathematical Sciences, University of Glasgow. These recruitment methods yielded of a pool of 156 interested volunteers. In the next section I describe how 48 subjects were chosen from this pool.

9.6.2 Selection

The name and email addresses of each subject were stored electronically. The list of subjects was divided based on volunteer gender (male 63.38%, female 36.62%). Subjects were sampled at random from these groups until 24 males and 24 females were chosen and notified through electronic mail. They were asked to visit a Web page containing an experimental timetable, select a small set of the most convenient times and respond via email. Experimental time slots were allocated based on subject preference and availability of suitable times. A time slot was allocated and a confirmation email sent.

Experimental subjects were assigned a unique experiment identifier in the range 101-148. This identifier was used during experimental data capture and analysis.

9.6.3 Subject Demographics and Search Experience

The average age of the subjects was 22.83 years (maximum 51, minimum 18, standard deviation = 5.23 years). Three quarters had a university diploma or a higher degree and 47.91% of subjects (23) had, or were pursuing, a qualification in a discipline related to Computing Science. The subjects were a mixture of students, researchers, academic staff and others. They had different levels of computing and search experience.

The subjects were divided into two groups – *inexperienced* and *experienced* – depending on their computing and search experience, how often they searched and the types of searches they performed. All were familiar with Web searching, and some with searching in other domains. The division of these groups was potentially problematic as subjects may not give an accurate account of their experience level. Table 9.1 shows the composition of each group and the differences between groups.

Table 9.1
Inexperienced and Experienced subject characteristics.

Factor	Inexperienced	Experienced
Number of subjects	24 (12 male, 12 female)	24 (12 male, 12 female)
Average search frequency	'Once or twice a week'	'Many times a day'
Use point-and-click interfaces	'Frequently' (3.58)	'A lot' (4.96)
Use Web search engines	'Frequently' (4.08)	'A lot' (4.92)

Subjects were asked to complete Likert scales asking how much experience they had with point-and-click interfaces, such as Microsoft Windows, and Web search engines. These results are reported in the last two rows of Table 9.1. The Likert scale values are in the range 1 to 5, where a higher value corresponds to more experience. The differences between subject groups were significant with a Mann-Whitney Test. [31]

Subjects were also asked to indicate which Web search engines they used and complete semantic differentials on how 'easy'/'difficult', 'stressful'/'relaxing', 'simple'/'complex' and 'satisfying'/'frustrating' the general use of these search engines was. This was potentially a good indicator of experience levels as I would expect subjects with more experience to be

[31] Experience with point-and-click interfaces, $U(24) = 441$, $p < .001$, experience with Web search engines, $U(24) = 396$, $p = .013$.

more competent searchers. Table 9.2 showed the average differential responses and the significance of the differences between subject groups with a Mann-Whitney Test.

Table 9.2
Search engine use (scale from 1 to 5, lower = better).

Differential	Inexperienced	Experienced	Significance[α]
easy	2.29	1.50	.004
relaxing	2.63	2.46	.475
simple	2.13	1.63	.045
satisfying	2.46	2.46	.156

[α] with a Mann-Whitney Test, $U(24)$.

The results show that those subjects classified as 'experienced' found using Web search engines significantly easier than the inexperienced group; to a certain extent this validated the subject classification. In the next section I describe the search tasks given to experimental subjects.

9.7 Tasks

In this section I discuss the search tasks attempted by experimental subjects. Tasks were divided into three categories and within these categories into six search topics. The tasks were designed to encourage naturalistic search behaviour by experimental subjects. I wanted subjects to interact with the experimental systems as though they were performing their own search. To do this, the tasks were placed within simulated situations as proposed in Borlund (Borlund, 2000b; 2000a). The technique asserts that searchers should be given search scenarios that reflect and promote a real information seeking situation. Figure 9.2 shows an example simulated situation.

Simulated Situation

Simulated work task situation: After your graduation you will be looking for a job in industry. You want information to help you focus your future job seeking. You know it pays to know the market. You would like to find some information about employment patterns in industry and what kind of qualifications employers will be looking for from future employees.

Indicative request: Find for instance something about future employment trends in industry, i.e., areas of growth and decline

Figure 9.2. Simulated situation taken from Borlund (2000a).

Simulated situations can be composed of two parts: the simulated work task situation and an indicative request. The simulated work task situation is a short 'cover-story' designed to provide context for a search. The indicative request is an indication, rather than an instruction, of how a search may be initiated. Previous studies have shown that the indicative request is not required for the simulated situation to engage the subject in the search and to promote natural searching behaviour on the part of the subject (Borlund, 2000a).

The simulated situations, such as that shown in Figure 9.2, are intended to achieve two main objectives. First, they promote a simulated information need in a subject. That is, the simulated situation should engage the subjects in the search by the identification of the searcher within the situation. As in Pilot Test 1, I offer subjects a choice of search tasks to go some way to ensuring they choose tasks of interest to them and can identify with the topic of the search. In Pilot Test 2, these tasks were tested for differences in their difficulty; no differences were found.

Second, the simulated situations position the search within a realistic context. The situation allows the experimental subject to provide his or her own interpretation of what information is required and allows them to develop the information need naturally. They permit a dynamic interpretation of relevance by experimental subjects. In forthcoming sections I describe the task categorisation and the search topics.

9.7.1 Task Categories

The tasks in this experiment were divided into three categories. Tasks were categorised based on their complexity and tried to encourage different types of information seeking behaviour. The aim of this approach was to create different types of needs to see how well the experimental systems performed for these differing types and to hopefully elicit different subject behaviours. The six stage Information Search Process (ISP) model (Kuhlthau, 1991) forms the basis of the task selection. I do not choose six task categories that correspond with the six stages in the ISP, but instead to the three types of searcher interaction that the model predicts; *background seeking*, *relevant seeking* and *relevant and focused seeking*. Through varying their complexity, this categorisation at least aims to encourage the types of interaction I would expect to see at each stage, in the hope that it may give a handle on what aspects of the search process each experimental system supports well, and what parts they do not. In earlier work (White *et al.*, 2003b) I proposed four categories of Web search; *fact search*, *decision search*, *search for a number of items* and *background search*. In an earlier study, Byström and Järvelin (1995) describe five task categories based on their complexity and *a*

priori determinability. The *a priori* determinability measures how well the searcher can determine the required task inputs (information necessary for their search), processes (how to find the required information) and outcomes (how to recognise the required information) based on the initial task statement. Through increasing the uncertainty associated with each of these factors an experimenter can control the complexity of the task. Table 9.3 shows the relationship between the ISP categorisation used in the experiment and this related work.

Table 9.3
Task categorisation and related work.

Related Work	Task category		
	Pre-focus	Focus formation	Post-focus
Information seeking behaviour (Kuhlthau, 1991)	*background*	*relevant*	*relevant or focused*
Task type (White *et al.*, 2003b)	*background*	*decision*	*fact* and *search for a number of items*
Task complexity (Byström and Järvelin, 1995)	*known, genuine decision task* and *genuine decision task*	*normal decision task*	*normal information processing task* and *automatic information processing task*

To create the *pre-focus*, *focus formation* and *post-focus* task categories I varied the number of potential information sources and type of information required to complete a task (Bell and Ruthven, 2004). Six search topics were chosen for the experiment and a pre-focus, focus formation and post-focus version of each category was created. In the next section I describe these topics.

9.7.2 Search Topics

Six search topics were tested in Pilot Test 2 and used in this experiment. The topics were chosen to be of general interest to participants and reflect searches they may be likely to perform. The simulated work task situations used in this experiment were tailored towards the information environment and the group of test persons. Borlund (2003) recommends that this tailoring is to include:

i. A situation which the test persons can relate to and in which they can identify themselves;
ii. A situation that the test persons find topically interesting, and;
iii. A situation that provides enough imaginative context in order for the test persons to be able to relate and apply the situation.

Tailoring of simulated work task situations is important in order to gain a trustworthy behaviour and IR interaction from experimental subjects. Table 9.4 shows the topic titles for the six search topics used.

Table 9.4
Titles of search topics used during experiment.

1. Applying to university	4. Third generation phones
2. Allergies in the workplace	5. Internet music piracy
3. Art galleries in Rome	6. Petrol prices

For each of these topics three search tasks were created to match the *pre-focus*, *focus formation* and *post-focus* task categorisation. Subjects chose one pre-focus, one focus formation, and one-post focus task. They choose tasks from a different search topic each time and were not allowed to choose more than one task for a particular topic. This minimised task learning effects. The search tasks are included in Appendix F.3, where Task A is the high-complexity 'pre-focus' task, Task B is the moderate complexity 'focus formation' task and Task C is the low complexity 'post-focus' task. In the next section I describe how tasks were allocated to subjects.

9.7.3 Task Allocation

Borlund (2000a) conducted a feasibility test and revealed a 'significant pattern of behaviour' amongst experimental subjects in the way they carried out the relevance assessments of the retrieved documents when using simulated work task situations. For this reason an experimental design was used that could reduce the likelihood that the use of one system or attempting one task, influenced the next task-system variation. A Graeco-Latin square design was used (Tague-Sutcliffe, 1992), that rotated both experimental systems and tasks.

Table 9.5 shows the experimental design. The factors in the table are the *tasks categories* ($T_{A\text{-}C}$) and the *experimental systems* (S_{Check}, S_{Recomm}, S_{Auto}).

Table 9.5
Graeco-Latin square experimental block design.

Subject	System/Task order		
	1	2	3
1	S_{Check} , T_A	S_{Recomm} , T_B	S_{Auto} , T_C
2	S_{Auto} , T_B	S_{Check} , T_C	S_{Recomm} , T_A
3	S_{Recomm} , T_C	S_{Auto} , T_A	S_{Check} , T_B

This square represents a block of subjects. There are 16 similar blocks of three subjects in the experiment (i.e., $16 \times 3 = 48$). In the next section I describe the experimental procedure.

9.8 Procedure

Each subject was asked to attempt each of the search tasks they had chosen. The order in which topics were presented, and the choice of which system a subject used for each search, was determined by the randomised experimental matrix given in the previous section. Experiments lasted between one-and-a-half and two hours, dependent on the amount of time required to complete questionnaires. Subjects were provided with light refreshments and were offered a five minute break after the first hour.

For each experiment the following steps were followed:

i. Subjects were welcomed and asked to read the introduction to the experiment provided on an 'Information Sheet' (Appendix F.1). This set of instructions was developed to ensure that each subject received precisely the same information. Subjects could retain the information sheet after the experiment.
ii. Subjects were then asked to sign two copies of a consent form, one for my attention, and one on the reverse of the 'Information Sheet', for the subject to keep.
iii. Subjects were then asked to complete an 'Entry' questionnaire (Appendix F.2). This elicited background information on the subject's education, previous general search experience, computer use experience and Web search experience.
iv. Subjects were given a tutorial on all experimental systems, followed by a training topic. The training topic was the same for all subjects and is included in Appendix F.3. This training topic gave subjects a chance to familiarise themselves with the interface components of the experimental systems. More details on subject training are given in Section 9.9.
v. Once comfortable with the training system subjects were given the first task sheet and asked to select one search task from the six in the allotted task category. No guidelines were given to subjects about the criteria to use when choosing a task.
vi. After selecting the task, subjects were asked to perform the search it required. They were given 15 minutes to search and could stop early if they were unable to find any more relevant information.
vii. After completing the search (either successfully or otherwise), the subject was asked to complete the 'Search' questionnaire (Appendix F.2).
viii. The remaining task sheets were given to subject, following steps v. – vii. Since the search topics were the same on all three task sheets subjects were not allowed to choose

the same topic as attempted in a previous search. Subjects were offered a five minute break after the first task (around halfway through the experiment).

ix. At the end of the experiment, the subject was asked to complete the post-experiment 'Exit' questionnaire (Appendix F.2) and an informal post-experiment interview was conducted.

The 'Search' and 'Exit' questionnaires were designed based on the research questions that motivated the experiment, described in Section 9.12. In the next section I provide more details on how experimental subjects were trained.

9.9 Training

Since the experimental systems were unfamiliar to subjects, they received pre-search training on how to use them. A short time, around 30 minutes was allocated for training at the start of the experiment. The training session was broken down into a series of stages:

i. I explained the purpose of the systems i.e., that they all tried to improve the quality of the subject's query and some tried to select new search decisions on the subject's behalf.
ii. Subjects were introduced to the search interface components that appeared in all systems (e.g., top-ranking titles, pop-up summaries). I used printed screenshots of each of the three experimental systems to help describe these interface components.
iii. I gave subjects a live demonstration of each system using the same search query, 'information'.
iv. A training task (Appendix F.3) was issued and subjects were given the chance to attempt this task on a training system with no feedback (similar to Koenemann and Belkin (1996)). The training task gave subjects an opportunity to use the system in a realistic information seeking context and become accustomed to the interface features.
v. The training session stopped once subjects felt comfortable using the systems.

Subjects were allowed to comment or ask questions at any point during the session. Due to the large number of experimental participants and the relatively short duration of the experiment, 30 minutes was the maximum time afforded to each subject. In all cases this appeared sufficient for subjects to familiarise themselves with the systems.

9.10 Questionnaires

Questionnaires were the main method used to elicit subject opinion during the experiment. The questionnaires were typically divided up into a series of sections that contained questions on the same aspect of the search (e.g., 'Search Process', 'Interface Support'). To help the subject complete the questions, some introductory text was given at the start of each section. Figure 9.3 gives an example of such text from the 'Search' questionnaire.

> **Relevance Assessment**
> The Automatic and Interactive systems assumed that much of the information you viewed was relevant. In the Checkbox system you explicitly marked relevant items.

Figure 9.3. Example introductory sentence (taken from 'Search' questionnaire).

Three questionnaires were developed and distributed to experimental subjects at various points in the search: 'Entry', 'Search' and 'Exit'. These questionnaires are included in Appendix F.2 and contained three styles of question; *Likert scales*, *semantic differentials* and *open-ended questions*. In this section each style is explained and examples provided.

9.10.1 Likert Scales

The Likert scaling technique presents a set of attitude statements. Subjects are asked to express agreement or disagreement on a five-point scale. [32] Each degree of agreement is given a numerical value from one to five. A total numerical value can be calculated from all the responses received. Figure 9.4 shows an example Likert scale taken from the 'Entry' questionnaire.

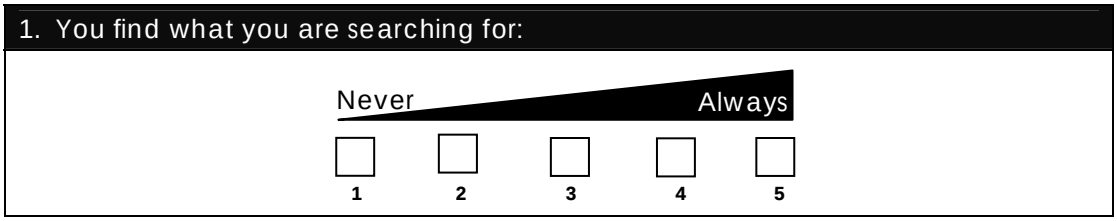


Figure 9.4. Example Likert scale (taken from 'Entry' questionnaire).

Likert scales are designed to show a differentiation among respondents who have a variety of opinions about an *attitude object* (i.e., anything that the subject may find good or bad), in this case how often they find what they are searching for.

[32] A five-point scale was preferred to seven or nine point scales as it made the analysis of subject opinion simpler and allowed trends in the results to be more easily identified.

9.10.2 Semantic Differentials

Another type of structured question is one that provides pairs of antonyms and synonyms, together with five-step rating scales. The word pairs refer to an attitude object, and respondents are asked to check one of the positions on each continuum between the most positive and negative terms. This type of scale is called a *semantic differential*. Figure 9.5 exemplifies a set of four semantic differentials.

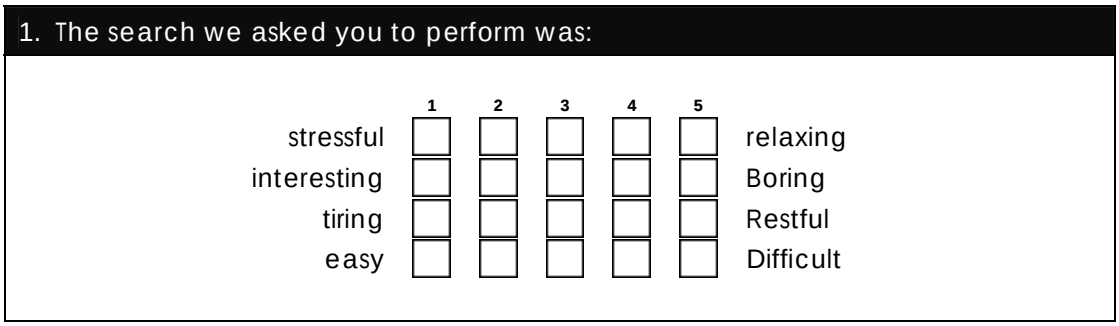


Figure 9.5. Example set of semantic differentials (taken from 'Search' questionnaire).

In this example, as in all differentials in the experimental questionnaires, the positive and negative terms are reversed in consecutive attitude objects. This ensures that subject attention does not waver when completing the questionnaires.

9.10.3 Unstructured Questions

In unstructured questions subjects were given the chance to freely reply without having to select one of several provided responses; these questions can be described as 'open-ended'. They are useful for revealing reasons why subjects feel the way they do and giving them a chance to comment freely on aspects of the system, the task or the experiment in general.

Subjects were issued with an 'Information Sheet' at the start of the search that showed them completed examples of Likert scales and semantic differentials. It was assumed that subjects would not need instructions on answering unstructured questions.

During the experiment, system logging recorded search activity at the interfaces to the experimental systems. In the next section I describe the logging procedure used.

9.11 System Logging

Log files were named based on the subject's unique identifier, the system and task attempted. The log file contains a header, which is written before any interaction. This contained the subject identifier, the task being attempted, the experimental system being used and the date and time of the experiment. Prior to starting the each search task I created this header using a small Java application. The interface to this application is shown in Figure 9.6. It was not important that this interface was intelligible to experimental subjects as only I used it. The buttons S1, S2 and S3 can be used to clear system log files, the 'id' boxes contain the subject identifier and the order in which systems are used. In Figure 9.6, subject 141 is using S2 then S3 then S1. The search topic (ST) boxes contain the identifier of the search category/topic attempted (e.g., A4 is the fourth topic on the high complexity task sheet).

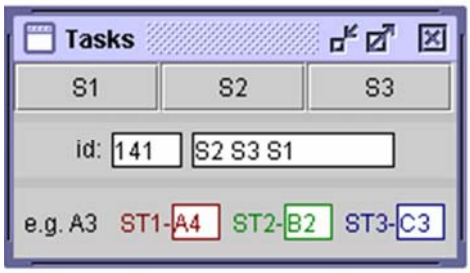


Figure 9.6. Java application for log header construction.

All searcher interaction with the experimental systems was also logged as a '<event> <timestamp>' pair and the timestamp was written as the number of milliseconds elapsed from midnight, January 1, 1970. This is a Java default and allowed times to be easily parsed and compared. Details of the tags used to denote the events and an excerpt from the log files are included in Appendix G.

The location of the mouse pointer is also logged every 0.25 seconds, and the locations of any mouse clicks are also recorded. From this log data I can analyse which parts of the interface subjects interact with and where they spend the most time. System usage data of this nature is useful for tracking exactly how subjects interact with these systems. In the next section I describe the experimental hypotheses tested during this experiment.

9.12 Hypotheses

The purpose of this experiment is to investigate the effectiveness of different forms of interface support for facilitating the use of relevance feedback in interactive search environments and the probabilistic framework described in Chapter Seven. The framework is

used to modify queries and select retrieval strategies based on relevance feedback provided by the searcher. This feedback can be *implicit* (inferred by the system from interaction) or *explicit* (provided intentionally to the system by the searcher); different experimental systems offer different ways of indicating what information is relevant.

This experiment investigates which form of interface support searchers prefer, the ability of the probabilistic framework to choose worthwhile terms and the appropriateness of the new retrieval strategies chosen or recommended. In this section the experimental hypotheses are described. These are:

Interface support (Hypothesis 1)

Subjects like the interface support provided by the experimental systems and find that it facilitates effective information access.

Information need detection (Hypothesis 2)

Subjects find the terms chosen by the probabilistic implicit feedback framework valuable and worthwhile.

Information need tracking (Hypothesis 3)

Subjects find the retrieval strategies chosen by the probabilistic implicit feedback framework valuable and worthwhile.

The hypotheses are analysed in three ways. The first examines the subjects' overall search behaviour; this analysis looks for changes in how subjects searched on the experimental systems. The second examines the search effectiveness of the three systems; on which system did the subjects have a most effective search? Finally I shall examine the subjects' perceptions of the three systems; did the subjects prefer one system over the others?

9.13 Sub-hypotheses

It is possible to divide the experimental hypotheses provided in the previous section into a number of sub-hypotheses to make the capture and analysis of data more straightforward. In this section each set of sub-hypotheses are described.

9.13.1 Hypothesis 1: Interface Support

Five aspects of the interface support offered by the experimental systems were tested in this experiment:

Relevance Paths and Content (Hypothesis 1.1)

Subjects find the information presented at the interface useful.

Term selection (Hypothesis 1.2)

Subjects want control in formulating new queries.

Retrieval strategy selection (Hypothesis 1.3)

Subjects want control in making search decisions.

Relevance assessment (Hypothesis 1.4)

Subjects want the experimental system to infer relevance from their interaction.

Notification (Hypothesis 1.5)

Subjects find system notifications helpful and unobtrusive.

9.13.2 Hypothesis 2: Information Need Detection

This hypothesis assesses the effectiveness of the information need detection part of the probabilistic framework. To test it, subject opinion on the terms chosen by the term selection model was elicited. I divide the hypothesis into two sub-hypotheses based on their *value* (can be helpful during a search) and *worth* (is correct and accurate).

Value (Hypothesis 2.1)

Query modification terms chosen by the framework are relevant and useful.

Worth (Hypothesis 2.2)

Query modification terms chosen by the framework approximate subject information needs.

9.13.3 Hypothesis 3: Information Need Tracking

The information need tracking component of the system looked for changes in the information needs of searchers as they searched. The information need tracking component is tested via subject perceptions of the retrieval strategy selected by the system. That is, the component is evaluated through subject perceptions of the resultant *search strategy*, not the perceived extent of the change. There are two sub-hypotheses that, in a similar way to Hypothesis 2, are based on the value and worth of the component:

Value (Hypothesis 3.1)

The retrieval strategies chosen by the framework are beneficial.

Worth (Hypothesis 3.2)

The retrieval strategies chosen by the framework approximate changes in the information needs of subjects.

9.14 Chapter Summary

In this chapter the methodology has been presented for a user experiment to: (i) investigate interface support mechanisms to assist users of information retrieval systems and (ii) evaluate the effectiveness of the probabilistic implicit feedback framework in realistic search environments. The hypotheses for the experiment have been introduced and the document domain, tasks, subjects and experimental procedure have been described. In this chapter, the experimental systems used to test the hypotheses were briefly introduced. In the next chapter these systems are described in more detail.

Chapter 10

Experimental Systems

10.1 Introduction

Three experimental systems were created to test the hypotheses proposed in the previous chapter. The systems vary subject control over three main classes of decisions that users of such systems must make: selecting query terms, indicating relevance and making new search decisions. The experimental systems were: (i) a system that allowed subjects to directly communicate what information was relevant, provided support in creating new queries and allowed searchers to decide how these queries were used, (ii) a system that gathered relevance indications through implicit feedback, recommended new queries and made recommendations on how these queries should be used, and (iii) a system that used implicit feedback, automatically refined the query and made search decisions on query use on the subject's behalf. Each system offers different types of interface support, and where appropriate uses the techniques described in Chapter Seven. In this chapter I describe the experimental systems, their similarities and their differences.

10.2 Overview of Systems

The systems developed were interfaces to Web search engines that provided added support in creating search queries and making search decisions (i.e., re-searching the Web, reordering document lists and reordering lists of Top-Ranking Sentences). The names given to the systems during the experiments were based on their distinguishing features. The three experimental systems and *search activities* on each were:

i. **Checkbox:** searchers control relevance indication and query generation; searchers control query word selection; searchers control query execution.

ii. **Recommendation:** searchers delegate relevance indications and query generation; searchers control query word selection; searchers control query execution.

iii. **Automatic:** searchers delegate relevance indication and query generation; searchers delegate or control query word selection; searchers delegate or control query execution.

A summary of the responsibilities for all search activities is given in Table 10.1.

Table 10.1
System and subject responsibilities for search activities.

	System		
Search Activity	Checkbox	Recommendation	Automatic
Query Modification	System and Subject	System and Subject	System
Relevance Indication	Explicit	Implicit	Implicit
Retrieval Strategy Selection	Subject	System and Subject	System

The role of the subject in query modification is different in the Checkbox and Recommendation systems. In the Recommendation system they choose additional terms from those *recommended*; if a term is irrelevant subjects can ignore it. The Checkbox system selects additional terms and appends these to the original query in an editable text box. The subject is then responsible for retaining or removing terms to formulate the new query; if a term is irrelevant searchers have to delete it.

The experimental systems share a number of underlying features and differ in those necessary to test the research hypotheses outlined in the previous chapter. The aim of this thesis was not to develop an optimal search interface. The interfaces I constructed were developed for experimental purposes and were sufficient to allow an investigation of implicit feedback and interface support mechanisms. The probabilistic framework described in Chapter Seven is used by all systems to make decisions about query terms and, in the Recommendation and Automatic systems, to select retrieval strategies. In the next section the similarities and differences between the experimental systems are described.

10.3 Similarities and Differences

The systems share many features and differ in only a few. The differences between systems are limited to those necessary to test the research hypotheses.

10.3.1 Similarities

In this section the system features common to all three systems are described. Among other things, the systems share the same architecture for retrieving documents and selecting Top-

Ranking Sentences, general interface components, term selection model and method for scoring sentences and documents.

10.3.1.1 Retrieval Architecture

The same retrieval architecture underlies each of the three systems and is described in Chapter Three. All systems are implemented in Dynamic HTML (DHTML) and the client-side code for all systems is written in JavaScript. A submitted query is passed to the Google commercial Web search engine and the top-ranked documents are retrieved and the Top-Ranking Sentences selected. Google was chosen for the size of its index, the frequency with which this index is updated and the existence of a Java Application Programming Interface that allowed me to easily query the search engine. [33] The best sentences from all top-ranked documents are used to construct a list of Top-Ranking Sentences, presented to the searcher at the interface. A term space containing all unique terms in the most relevant documents is also constructed. [34] This space is used by the Jeffrey's Conditioning Model; each term in the space is considered a candidate for query modification.

10.3.1.2 Interface Components

The interfaces to the experimental systems in this experiment used titles, summaries and sentences as described in Chapter Five and in Pilot Test 1. However, unlike the interfaces used in Pilot Test 1 these interfaces use mouse clicks *on* search results rather than movements *over* search results as an indication of the relevance. Clicks show the subject the next step in the relevance path or open Web documents. Since the subject must act 'explicitly' (although not for the purpose of communicating relevance) each of these actions are assumed to be more reliable indicators of subject interests than mouse movements. A click represents a conscious effort by the subject and a break in their cognitive processes; clicks are normally intentional and can therefore be more reliable implicit relevance indicators than mouseovers. With mouseovers it can be difficult to determine what actions are intentional and which are accidental, arising through the movement of the mouse to another part of the screen. To follow a relevance path subjects must 'hover' over representations for a short period of time and click arrows next to representations as shown in Figure 10.1.

[33] http://www.google.com/apis/

[34] In the same way as Chapter Four, query-relevant Top-Ranking Sentences were selected from the top 30 retrieved documents to ensure the systems responded to the subject in a timely manner.

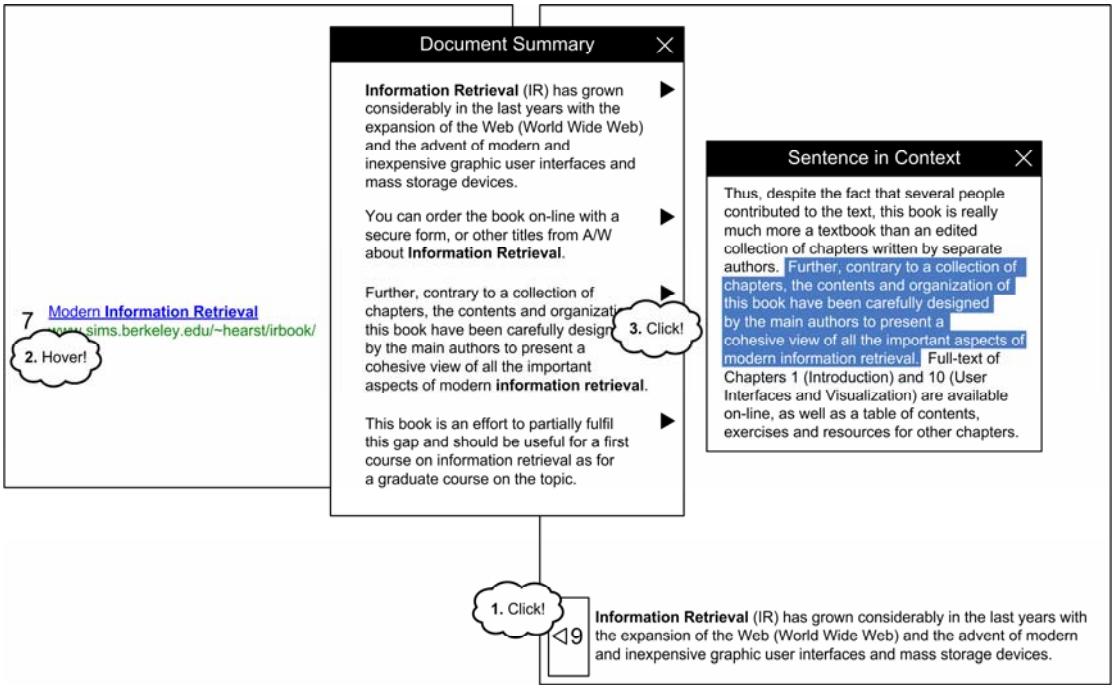


Figure 10.1. Necessary actions for relevance path traversal.

Subjects can visit the source document of any document representation by clicking its textual content. To see the next step in the relevance path they must *click* arrows next to representations (e.g., click arrow next to top-ranking sentence to highlight source document) or *hover* over representations (e.g., hover over title to see summary).

Since the Recommendation and Automatic systems used the movement of the mouse pointer over parts of the interface as an indication of relevance a timing mechanism was implemented to ensure these 'hovers' were intentional. That is, a searcher would have to remain over a document title for two seconds before the pop-up summary window appeared. In the studies in Chapter Four I demonstrated that a timing mechanism can be useful to tackle problems caused by accidental mouseovers in feedback systems that use implicit feedback techniques. Also, when the document summary appears, the other information in the background of the interface darkens and is disabled to ensure that it does not interfere with the examination of the summary and cannot be clicked accidentally.

The Recommendation and Automatic systems used the information that subjects interacted with as implicit feedback of their interests. The systems used this feedback to build a richer body of evidence and choose query terms to represent the information interacted with. In Table 10.2 I show the actions necessary for these systems to identify what is of interest to searchers; the indications in bold are those that comprise a relevance path. Providing the

bolded indications in order, from top-ranking sentence to sentence in context means a searcher will traverse a complete relevance path.

Table 10.2
Implicit relevance indications.

Document Representation	Indication	Interpretation
Top-Ranking Sentence (TRS)	1. Click TRS	View document
	2. ***Click arrow on TRS***	Highlight document title
	3. Click '…' [35] at end of TRS	View remainder of sentence
Title	1. ***Hover for over two seconds***	View summary
	2. Click title	View document
Summary	1. Click text	View document
	2. ***Click arrow on Summary***	View sentence in context
Summary sentence	1. Click text	View document
	2. ***Click arrow on Summary***	View sentence in context
Sentence in context	1. ***Click text***	View document

A simple governing interaction model is that interacting with a document representation in any way is interpreted as a positive relevance indication. It can be seen in Table 10.2 that subjects can view the source document of a representation simply by clicking on its textual content. The mouse pointer changes when over these representations to indicate that they can be clicked. Also, all interaction with the document summary is regarded as an indication of interest. That is, all clicks in the summary are an indication of relevance for the text in the summary.

10.3.1.3 Term Selection Model

All systems use the term selection model chosen from the probabilistic implicit feedback framework to select query modification terms. As described later in this chapter they differ in how these terms are subsequently used.

10.3.1.4 Document/Sentence Reordering

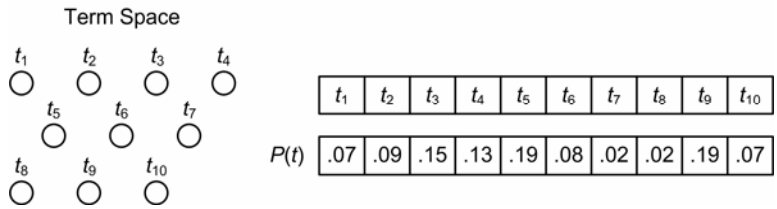
Two of the four possible retrieval strategies available for selection by the system or the subject involve reordering the most relevant documents and Top-Ranking Sentences. In my approach sentences are synonymous with small documents and the same approach is used to reorder documents and sentences. For consistency, in this section the term 'document' is synonymous with 'sentence'.

[35] To avoid unnecessary interface clutter, only the first 250 characters of a top-ranking sentence are shown at the interface. Ellipses are shown at the end of sentences where more text is available. Clicking on these ellipses shows the remainder of the sentence in a small area next to the mouse pointer. This is also used as an indication of interest.

The systems use a variation of the *tf.idf* approach to reorder the documents with respect to the query terms they contain. The inverse document frequency (*idf*) is regarded as a measure of importance of the term in the collection. In the approach used here, the values of the $P(t)$ assigned to terms in the term space can also be regarded as a measure of importance and the values are used instead of *idf* in this reordering. Unlike *idf* values, the $P(t)$ values alter to reflect the changing importance of the terms during a search. I now present an example of how this approach is used to rank documents or sentences.

Example 10.1: Document Reordering

In this example there are five documents (D_1, D_2, D_3, D_4 and D_5) and the term space is in the same state as at the end of Example 7.1 (Chapter Seven). There are ten terms in the term space and the query contains terms t_2, t_5, t_8 and t_9. The weights assigned to each term in the term space are:



These weights are not revised during the reordering, but may change during the search, as a result of searcher interaction. They are used in conjunction with the frequency of terms within documents to produce a retrieval status value (RSV) used to rank documents. The term frequencies for each of the five documents in this example are:

D_1	t_1	t_3	t_5
	3	4	1

D_2	t_5	t_9
	6	4

D_3	t_2	t_7	t_8	t_9
	3	1	1	4

D_4	t_2	t_5	t_8
	6	1	3

D_5	t_4	t_6
	7	1

The documents are then ranked based on the scores of terms that reside in them and queries:

$$\boldsymbol{D_1} \quad \{t_2, t_5, t_8, t_9\} \cap \{t_1, t_3, t_5\} = \{t_5\}$$

$$\Rightarrow (.15 \times 1) = \mathbf{0.15}$$

$$\boldsymbol{D_2} \quad \{t_2, t_5, t_8, t_9\} \cap \{t_5, t_9\} = \{t_5, t_9\}$$

$\Rightarrow (.19 \times 6) + (.19 \times 4) = \mathbf{1.90}$

$\boldsymbol{D_3}$ $\{t_2, t_5, t_8, t_9\} \cap \{t_2, t_7, t_8, t_9\} = \{t_2, t_8, t_9\}$

$\Rightarrow (.09 \times 3) + (.02 \times 1) + (.19 \times 4) = \mathbf{1.05}$

$\boldsymbol{D_4}$ $\{t_2, t_5, t_8, t_9\} \cap \{t_2, t_5, t_8\} = \{t_2, t_5, t_8\}$

$\Rightarrow (.09 \times 6) + (.19 \times 1) + (.19 \times 3) = \mathbf{1.30}$

$\boldsymbol{D_5}$ $\{t_2, t_5, t_8, t_9\} \cap \{t_4, t_6\} = \varnothing$

$\Rightarrow \mathbf{0}$

The document order based on the RSV is therefore D_2, D_4, D_3, D_1 and D_5. The documents that contained the query terms were ranked above those without. The top-ranked document (D_2) was ranked highly because it contained a large number of query terms that were regarded as important (i.e., had a high $P(t)$ value). It is conceivable that there could be a different document order if the searcher had interacted with different information before this action.

10.3.1.5 Initial Query Input and Restrictions on Length

The same initial query input screen is used by all experimental systems. This is the part of the system where the search typically begins. The look and feel of this initial interface is intentionally simple and contains a text input box, a submit button and access (through a link) to some details on the query syntax supported by the systems and the automatic exclusion of stopwords. Turtle (1994) found that searchers with no training in query formulation can experience difficulties in generating sound queries. He showed that unstructured queries containing only queries to separate the terms are more effective for searchers. Queries with embedded operators such as *, -, $ and + are meant to offer searchers greater control in query formulation. However, searchers may have difficulty using these operators because they are not consistent between search systems (Shneiderman *et al.*, 1997).

To prevent possible bias caused by previous search experience experimental subjects were not told that the systems were interfaces to Google. Queries were restricted to lists of terms separated by spaces and were automatically combined by the search engine. Queries submitted to Google have term order sensitivity (Muramatsu and Pratt, 2001). The system uses term proximity and exact phrase matching to give documents where terms that occur in the same order as the query and close proximity a higher weight. The concatenation of terms to form search phrases using “” was permitted. The use of search engine specific syntax such as ‘site:’ and ‘link:’ was discouraged.

Due to restrictions imposed by Google queries could not be longer than ten words. If the subject tried to submit a query of more than 10 words to any experimental system they were presented with an error message as shown in Figure 10.2.

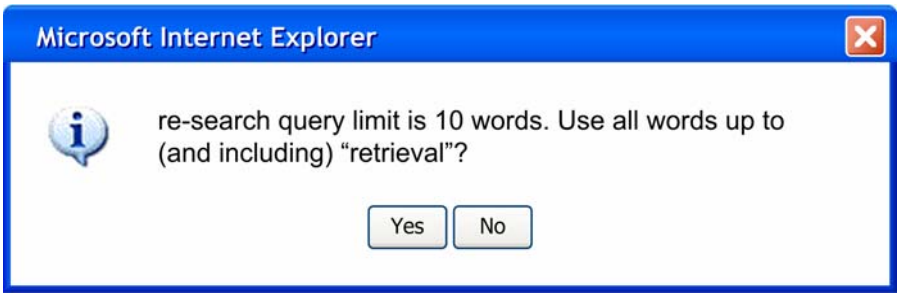


Figure 10.2. Query length notification message.

The query is truncated at the tenth word but before doing so the searcher is asked if they want to proceed. In Figure 10.2 the tenth word in the query is 'retrieval' and all words that follow this will be ignored by the search system.

10.3.1.6 Reversal of Retrieval Strategies

In his book 'Designing the User Interface', Shneiderman (1998) stresses the importance of allowing users to reverse the effects of their interaction. In each experimental system the subject has the option to reverse the effect of any search decision made by them or by the system. This is done using a clickable 'undo' button shown in Figure 10.3.

Figure 10.3. Retrieval strategy reversal ('undo') button.

The button intentionally resembles the 'back' button in Internet Explorer, the browser used for these experiments. Although the functionally was different (i.e., it does not take subjects back to the previous Web document), the underlying intent is similar (i.e., to reverse the last action). I assume that clicking this button is an indication of dissatisfaction with the outcome last search decision. The underlying implicit feedback framework does not consider such negative feedback, only positive indications are used. However, it is plausible that the reversal of decisions and the traversal of short relevance paths (without visiting the source document) could be used as an indication of disinterest and to lessen the weight of terms in those representations and modify the decision boundaries used when selecting new retrieval strategies.

10.3.1.7 Notification of Actions

The Recommendation and Automatic systems select new query terms and make search decisions for the subject as they search. They notify them of this by displaying messages in the bottom left-hand corner of the interface. However, if the searcher is looking at information in a different part of the screen they may be unaware a retrieval strategy has occurred or been recommended to them. To be sure they notice these actions the systems place an 'idea bulb' next to the mouse pointer when a strategy is followed or a recommendation is made. This is shown in Figure 10.4.

Figure 10.4. The 'idea bulb' notification at appears next to the mouse pointer (pictured).

This bulb disappears when the subject interacts with the suggested terms or notification messages in any way. Since the mouse pointer is the primary means of interacting with the search interface, communicating decisions via the pointer notifies searchers, but does not intrude on their search (i.e., they can simply ignore the bulb). The idea bulb supplements the Recommendation and Automatic system notifications, which appear in one part of the interface and may not be immediately noticeable if searcher attention is elsewhere. In this section I have described the similarities between the three experimental systems. In the next section I outline the differences between the experimental systems.

10.3.2 Differences

The differences between systems were necessitated by the hypotheses tested in this experiment. More specifically, the systems vary subject control over three main classes of decisions: selecting query terms, indicating relevance and making new search decisions (i.e., choosing retrieval strategies). In this section I describe the differences between systems in a set of pair-wise comparisons.

10.3.2.1 Checkbox and Recommendation

There are three differences between these systems; how new queries are created, how search decisions are made and how relevance information is communicated. The Checkbox system awaits the searcher's instruction and selects query terms that describe the information the searcher has explicitly marked as relevant. The searcher can add or remove their query terms. The Recommendation system does not require such direct indications and presents a list of potentially useful terms that can be added to the initial query. In both systems the subject has

complete control over when a search decision is made and which decision is made. The Recommendation system recommends the retrieval strategy to the searcher, based on the estimated amount of information need change. The searcher has the option on whether to accept this recommendation.

10.3.2.2 Checkbox and Automatic

The differences between these systems lie in how search decisions are controlled and how relevance indications are provided. Retrieval strategies are controlled by the subject in the Checkbox system and by the information need tracking component in the Automatic system. Relevance is communicated directly (explicitly) in the Checkbox system and indirectly (implicitly) in the Automatic system.

10.3.2.3 Recommendation and Automatic

These experimental systems differ in how terms are selected for query modification and how search decisions are controlled. The Automatic system chooses terms and retrieval strategies on the subject's behalf. In contrast, the Recommendation system recommends terms and strategies.

Overall, the systems differ in the amount of control they offer to the searcher. With additional control there is also extra responsibility for making query modification decisions and choosing appropriate retrieval strategies. In the Checkbox system there is also the additional burden of explicitly marking document representations. Beaulieu and Jones (1998) showed that such additional control is not always preferred by searchers and places additional demands on their finite cognitive resources. However, these systems allow searchers to indicate what information has relevant properties and may be more accurate than systems without this burden. In the next section the experimental systems are described in more detail.

10.4 Systems

The experimental systems each consist of an interface connected to Google with the architecture defined in Section 10.3.1.1. In this section I describe each of the three systems.

10.4.1 Checkbox System

This system allows subjects to communicate directly which document representations are relevant. A checkbox is shown next to each representation and the subject can choose which representations to mark. Marking a representation is an indication that its contents are

relevant. The interface for this system is shown at two points during a search in Figure 10.6. The first part of the figure shows the summary window and sentence in context requested by the searcher. When 'Summary' or 'Sentence in Context' windows are requested the background darkens and is disabled to focus searcher attention on the active representation. Unlike the other experimental systems, all document representations in this system have checkboxes next to them that allow the searcher to mark them as relevant. In the second part of Figure 10.5 the searcher has requested assistance in creating a new query using the representations marked and extra terms have been added to the editable query entry box.

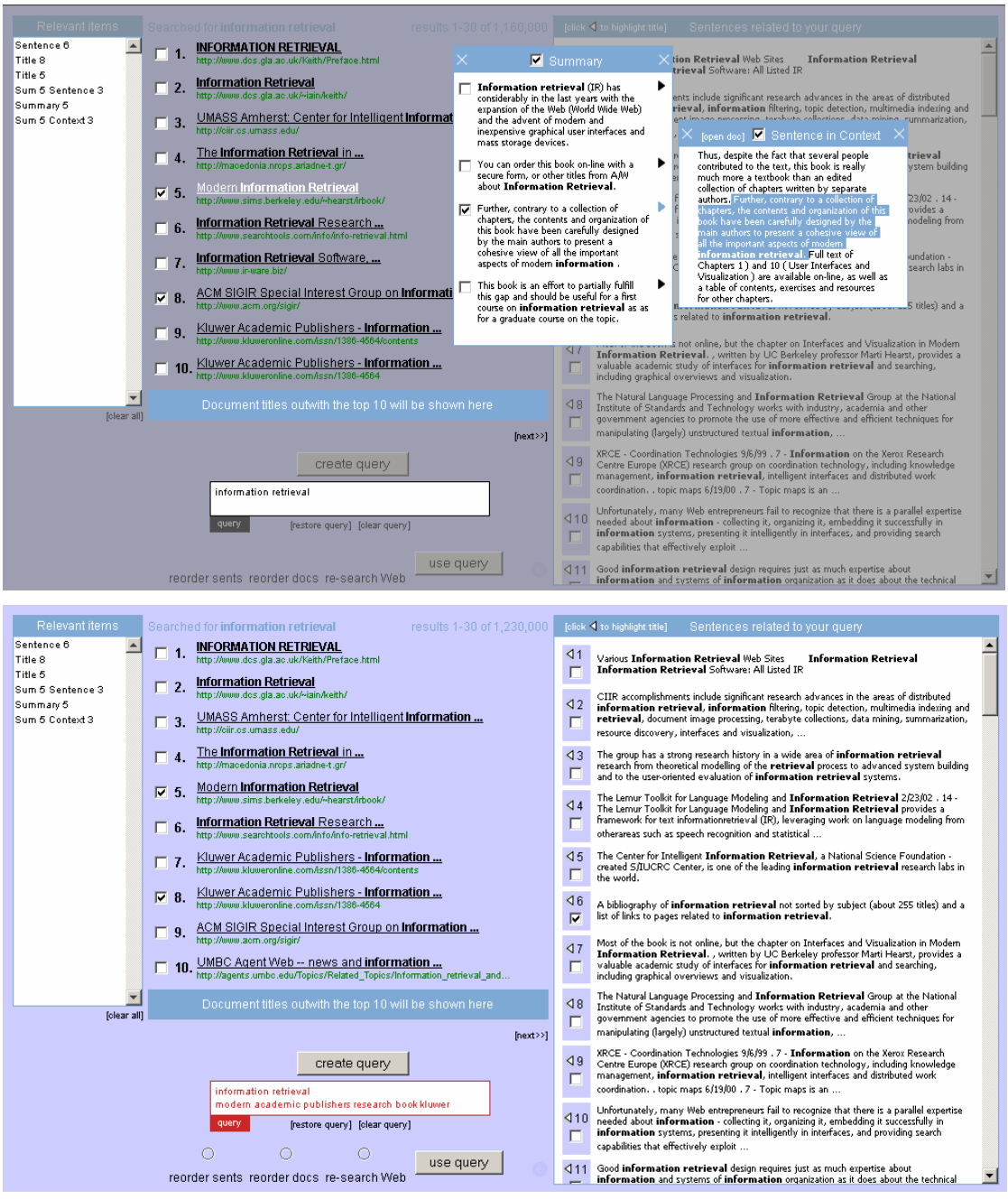


Figure 10.5. Checkbox system interfaces.

On the far left of the interface is a list of 'Relevant items' that describes which representations the subject has chosen so far. The nature of the interface, with pop-ups etc. is such that the subject may not see all representations they have marked relevant. This list allows them to keep track of what they have marked. In Figure 10.5 a number of document representations have been marked by the searcher. At any point the searcher can clear all representations they have marked or double-click an entry in the list of marked representations to highlight that particular representation. For clarity, from this point on all search interfaces are shown without the document summary and sentence in context pop-up.

The interface contains control options that allow the subject to request support with query formulation, modify the query and choose retrieval strategies. These options are shown in Figure 10.6.

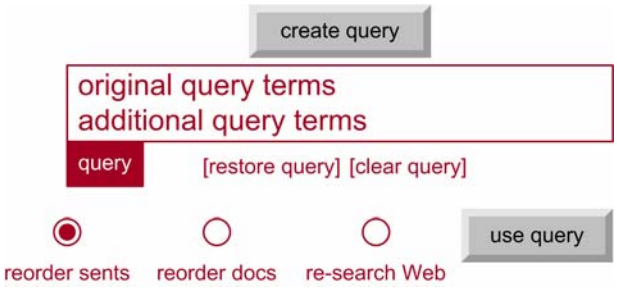


Figure 10.6. Term/retrieval strategy selection in the Checkbox system.

When they are satisfied with the document representations marked the subject can click the 'create query' button and a new query will be constructed. The presence of the button allows subjects to request assistance with query formulation. The term selection model treats each marked document representation as a separate relevance path and the order they were marked in is important. The terms chosen to expand the query are the six terms with the highest probability of relevance ($P(t)$ from Equation 7.10). These terms are appended onto the original query and presented in a search box for the searcher to edit, shown in Figure 10.6. The new query terms will be shown on a new line, below the original query.

In the Checkbox system the subject has control over the nature and timing of when search decisions are made. That is, at any time during their search they can choose the retrieval strategy (i.e., when to reorder the sentences, reorder the documents or re-search the Web) they feel is most appropriate.

10.4.2 Recommendation System

In the Recommendation system there are no checkboxes for the subject to explicitly mark what document representations are relevant. Instead, the system implicitly infers what is relevant from representations the subject has expressed an interest in through viewing or clicking. The search interface for the Recommendation system is shown in Figure 10.7.

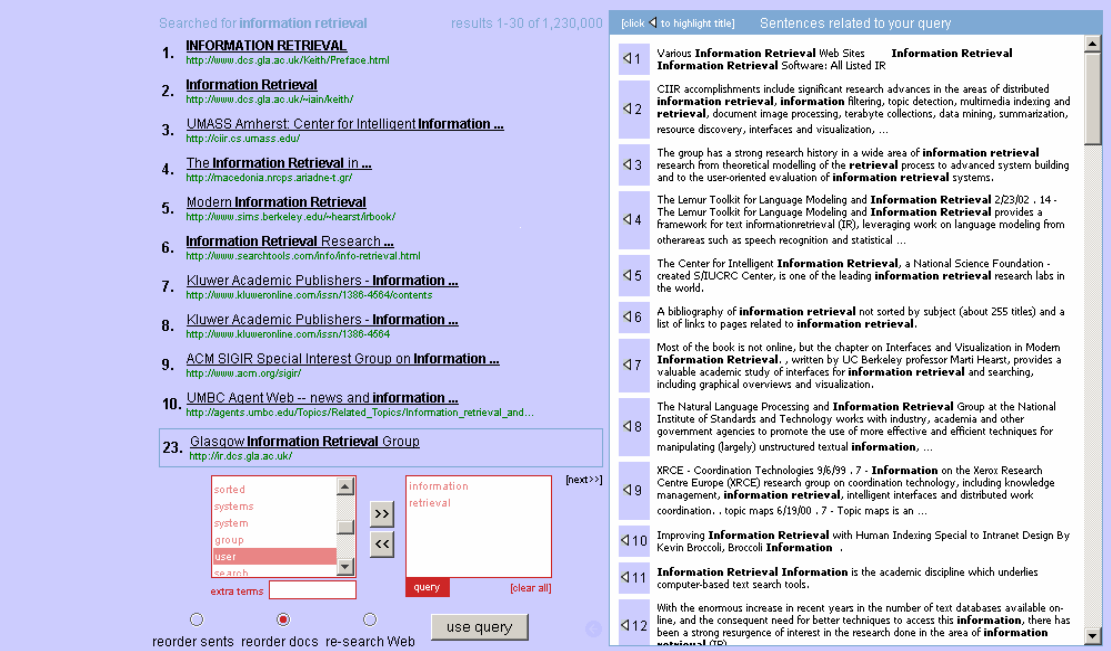


Figure 10.7. Recommendation system interface.

At intervals of five [36] relevance paths, the system chooses a new set of potentially useful query terms and a retrieval strategy based on the level of change in its internal information need formulation since the last subject-controlled query submission. Terms are chosen that reflect the information viewed. The degree of change since the last time a new result set was generated is used to select the action the system will perform. The system chooses the top 20 most relevant terms and presents these in the 'Recommended Terms' box (Figure 10.8).

[36] This was chosen in pilot testing (including Pilot Test 1) and allowed the system to build a body of evidence sufficient to make decisions.

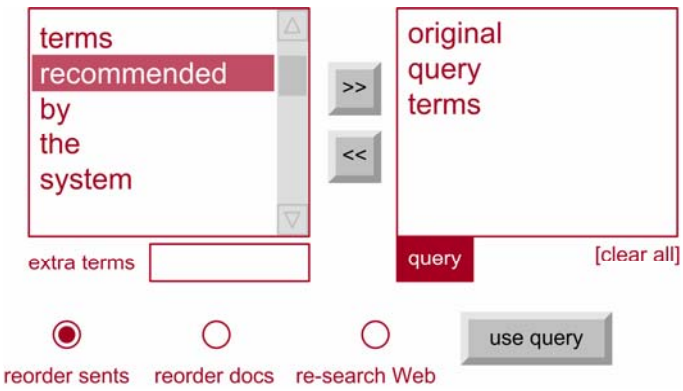


Figure 10.8. Term/retrieval strategy selection in the Recommendation system.

The subject can then control which terms are added to the query. Terms can also be deleted from the query. The '>>' and '<<' buttons can be used to transfer terms between the recommended list and the query. There is an 'extra terms' box where subjects can add additional terms to the query that are not in recommended terms list. When the subject clicks the '>>' button or presses 'enter' the term(s) in the box are added to the query. If the box contains more than one term the contents of it are tokenised and each token is added to the query separately. To reduce the number of erroneous terms that are transferred the searcher is only able to select and add one term at a time. Informal pilot testing of the interface revealed that subjects rarely want to add blocks of contiguous terms to the query at the same time. They preferred instead to be careful and selective about the terms they chose.

The system highlights the radio button for the retrieval strategy recommended by the experimental system. The subject does not have to agree with this recommendation and can choose another strategy or simply do nothing.

10.4.3 Automatic System

The Automatic system obtains its relevance assessments implicitly in the same way as the Recommendation system. However, the system retains control of the search decisions taken and the terms used. Rather than recommending terms and retrieval strategies, the Automatic system chooses them, without direct instruction. The interface is shown in Figure 10.9.

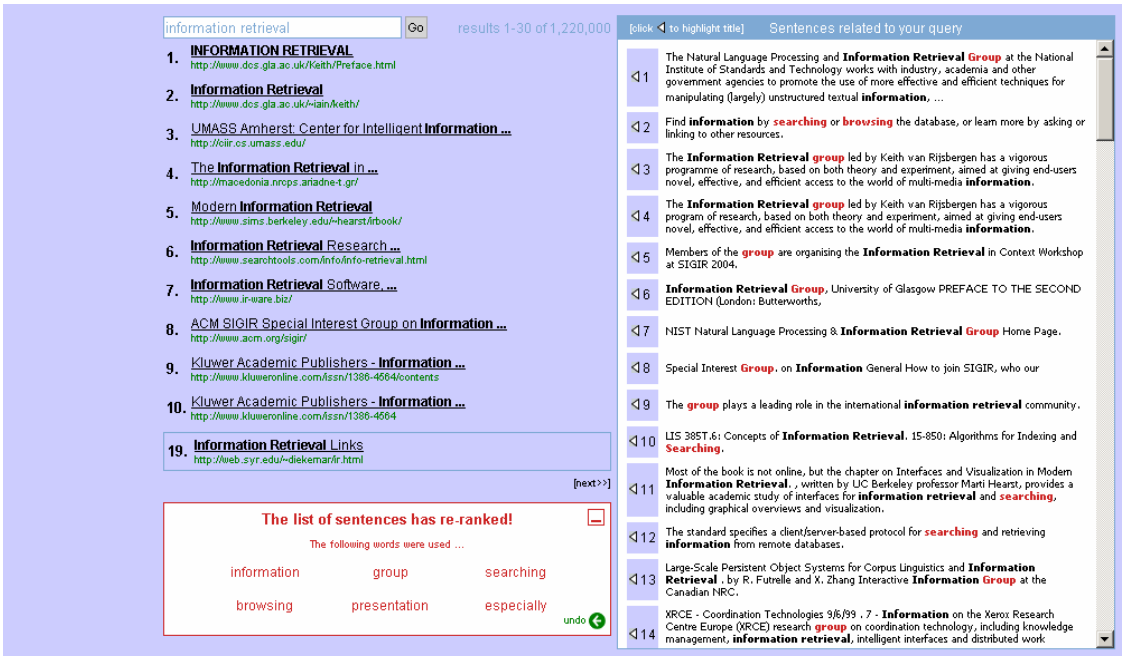


Figure 10.9. Automatic system interface (with maximised notification, Figure 10.10).

This system allows the subject to edit their original query and retrieve a new set of documents. No provision is made for the subject to formulate a query for reordering sentences or documents, these actions are controlled by the system. The system chose terms automatically and acts on the subject's behalf. Since subjects could not control the terms that were used it was necessary for this system to be able to replace the original query terms. If the information need changed during the search, the presence of the original terms would have meant the system could not totally adapt to that change. As in the Checkbox and Recommendation systems the new query is limited to a maximum of 10 terms as this is the maximum number of query terms supported by the Google search engine.

The system notified subjects that a new set of documents had been retrieved or the already retrieved information had been restructured using notifications at the search interface. These notifications were in two forms: maximised notification and minimised notification, shown in Figure 10.10 and Figure 10.11 respectively.

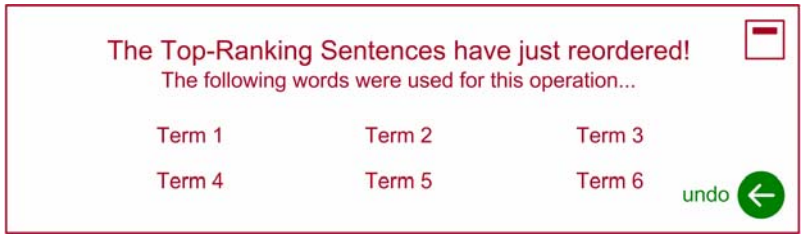


Figure 10.10. Maximised Automatic system notification.

Figure 10.11. Minimised Automatic system notification.

The minimised notification is less intrusive, but is also less informative and does not tell the subject which terms are used. The subject can switch between the different forms of notification by clicking on the notification message.

10.5 Chapter Summary

Three experimental systems have been described this chapter. These systems were created to test the hypotheses given in Chapter Nine. The systems allow relevance information to be communicated in different ways, and for subjects to have varying degrees of control over how new queries are created and how search decisions are made during their search. All systems use the probabilistic implicit feedback framework described in Chapter Seven. In the next chapter the results of the experiment involving these systems are presented and analysed.

Chapter 11

Experimental Results and Analysis

11.1 Introduction

In this chapter the results of the user experiment described in the two preceding chapters are presented. The experiment tests three search interfaces that vary searcher control over interface decisions, and the probabilistic implicit feedback framework (from Chapter Seven) that underlies them. Experimental subjects attempted search scenarios on the experimental systems and provided feedback on their experience through questionnaires and comments made during informal discussions. I focus on results that relate to each of the three research hypotheses originally proposed at the end of Chapter Nine:

Interface support (Hypothesis 1)

> The interface support provided by the experimental systems was liked by subjects and facilitated effective information access.

Information need detection (Hypothesis 2)

> Subjects found the terms chosen by the probabilistic implicit feedback framework valuable and worthwhile.

Information need tracking (Hypothesis 3)

> Subjects found the retrieval strategies chosen by the probabilistic implicit feedback framework valuable and worthwhile.

The hypotheses are tested in terms of search effectiveness and subject preference. A total of 48 subjects, with different levels of search experience participated in the experiment. Subjects were classified into two groups – *inexperienced* and *experienced* – each containing 24 volunteers and a mixture of males and females. Results are presented for inter-system (Checkbox *versus* Recommendation *versus* Automatic) and inter-group (inexperienced *versus*

experienced) comparisons. The significance of experimental results is tested at $p < .05$, unless otherwise stated. As in Chapter Ten S_{Check}, S_{Recomm} and S_{Auto} are used to denote the Checkbox, Recommendation and Automatic experimental systems respectively. In this chapter I also present results on the novel interface components (i.e., the relevance paths and increased information content at the search interface) and the search tasks.

The results presented in this chapter are based on questionnaire responses and system logs generated during interaction. The evidence is supported by informal subject feedback and my own observations. Questionnaires used five point Likert scales and semantic differentials with a lower score representing more agreement with the attitude object. The arrangement of positive (e.g., 'easy', 'relaxing') and negative (e.g., 'difficult', 'stressful') descriptors was randomised so that a positive assessment would be represented sometimes by a high score (i.e., approaching 5) and sometimes by a low one (i.e., approaching 1). This ensured that subjects applied due care and attention when completing the differentials (Busha and Harter, 1980). At the analysis stage the high positive scores are reversed so that in all cases the positive assessments were represented by low scores.

No assumptions are made about the normality of the data gathered during the experiment. Non-parametric statistical tests are used to test for statistical significance since these tests do not make any assumptions about the underlying distribution of the data. Also, since much of the data gathered was ordinal in nature (e.g., Likert scales and semantic differentials) these methods are more appropriate than their parametric equivalents. As described earlier, subjects were divided into two groups, *inexperienced* and *experienced*. The analysis presented involves *within-group* comparisons (e.g., one subject group with two or more systems) and *between-group* comparisons (e.g., comparing different subject groups on the same system). Where appropriate Dunn's *post hoc* tests (multiple comparison using rank sums) are applied to reduce the likelihood of Type I errors (i.e., rejecting null hypotheses that are true). The results across both subject groups are combined to form an 'Overall' group that gives a holistic view of the experimental findings for all subjects. The experimental design is a 2×3 factorial with *search experience* (2 levels) and the *experimental systems* (3 systems) as the main effects; tests are run for interaction between these where appropriate.

I begin this chapter by presenting results on the search process (Section 11.2) and the tasks attempted (Section 11.3). Tasks are analysed separately and relative to subject perceptions and measures of search effectiveness. This is followed by findings on the interface support (Hypothesis 1) (Section 11.4) and the terms and strategies selected by the probabilistic framework (Hypotheses 2 and 3) (Section 11.5 and 11.6 respectively). In Section 11.7 this

chapter concludes with a summary of the experimental findings. This experiment was in part a study of searcher control in interactive information retrieval. As such, the findings presented in this chapter focus on subjective impressions of the interface support mechanisms the experimental systems offer.

11.2 Search Process

In this section I present results on the search subjects performed. Whilst this analysis is not necessary to test the hypotheses, the factors may have an impact on subject perceptions. Each subject was asked to describe various aspects of their experience on each experimental system. The results presented are from questionnaire and informal subject comments, both during the search and after the experiment. Subjects were asked about their search and the quality of the information retrieved by each of the experimental systems.

11.2.1 Perceptions of Search

Subjects were asked to complete four semantic differentials about their search: 'relaxing'/'stressful', 'interesting'/'boring', 'restful'/'tiring' and 'easy'/'difficult'. The average value in relation to each positive differential is shown in Table 11.1. The 'Overall' value is derived from all four differentials and shows how the process is perceived across all subjects. For each differential in each subject group, the most positive average differential response is shown in bold. Below Table 11.1 and for each table in this chapter I use *n* to represent the number of trials in each cell. For example, in the table there are 24 trials in each 'Inexperienced' cell, 24 trials in each 'Experienced' cell and 48 trials in each 'Overall' cell.

Table 11.1
Subject perceptions of the search process (range 1-5, lower = better).

	Inexperienced			Experienced			Overall		
Differential	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
relaxing	2.75	2.33	**2.17**	2.67	2.25	**2.21**	2.71	**2.29**	2.19
interesting	2.70	2.54	**2.38**	2.08	**1.88**	2.21	2.40	**2.21**	2.30
restful	2.79	**2.71**	**2.71**	2.71	**2.25**	2.33	2.75	**2.48**	2.52
easy	2.75	**2.38**	2.67	2.58	**2.33**	2.50	2.67	**2.36**	2.59
all	2.75	2.49	2.48	2.51	2.18	2.31	2.63	2.34	2.40

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

A Friedman Rank Sum Test was run for each differential within each group. The test tries to answer the question: If the different systems really are identical, what is the chance that random sampling would result in sums of ranks as far apart (or more so) as observed? Since this analysis involved multiple comparisons, I use a Bonferroni correction to control the

experiment-wise error rate and set the *alpha level* (α) to .0125 i.e., .05 divided by 4, the number of tests performed. This correction reduces the number of Type I errors i.e., rejecting null hypotheses that were true. The results showed significant differences for the 'relaxing', 'interesting' and 'easy' differentials (*inexperienced*: all $\chi^2(2) \geq 14.26$, all $p < .001$) and 'relaxing', 'interesting', 'restful' and 'easy' differentials (*experienced*: all $\chi^2(2) \geq 14.83$, all $p < .001$ and *overall*: all $\chi^2(2) \geq 16.22$, all $p < .001$). [37] A Dunn's *post hoc* test was applied for each system in each subject group and found that for those differentials all differences were significant. The Recommendation system generally created a more pleasant search experience than the other systems; the Checkbox system was generally worse. Subjects found searches in the Recommendation system more interesting than in the other systems. The interface support provided by the system may have enabled subjects to view a broader range of documents or more fully explore those that interested them rather than dedicating time to explicitly assessing relevance.

The analysis also revealed significant differences in the differentials between the subject groups for the 'interesting', 'restful' and 'easy' differentials with a Mann-Whitney Test (all $U(24) \geq 399$, $\alpha = .0125$, all $p \leq .011$). To test for interaction effects between the two main effects; search experience and experimental system, and the dependent variable (i.e., the differential value) I ran a Kruskal-Wallis Test for each differential using the technique described by Meddis (1984, pp. 305-313). The test tries to answer the question: If the populations really have the same median, what is the chance that random sampling would result in sums of ranks as far apart (or more so) as observed in this experiment? The test returns an *H*-statistic that can use the Chi-square test to determine its significance (Siegel and Castellan, 1988). The results showed that for all differentials there was no significant interaction between search experience and system ($\chi^2(2) = 2.10$, $p = .35$). This demonstrates that the influence of the main effects on one another was not sufficient to affect the conclusions I can draw about each of them. This approach will be used where appropriate to test for interaction effects during this chapter.

11.2.2 Information value

The quality of information retrieved by search systems may have affected subject perceptions of them and could therefore influence the results described later in this chapter. To measure the quality of the information retrieved by the experimental systems throughout the search

[37] For large sample sizes the critical values of the Chi-squared distribution can be used to determine the statistical significance of the Friedman Rank Sum Test (Siegel and Castellan, 1988). Chi-Squared tests are represented by the notation χ^2(*degrees of freedom*).

subjects were asked for their opinion. On a Likert scale subjects indicated the extent to which they agreed with the attitude statement: *I think there was better information available (that the system did not help me find)*. The average responses, for different systems and different subject groups are shown in Table 11.2.

Table 11.2
Quality of information retrieved by the experimental systems (range 1-5, lower = better).

Inexperienced			Experienced			Overall		
S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
3.00	**2.92**	2.96	3.08	**3.04**	2.96	3.04	**2.98**	2.96

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

Subjects commented that they did not notice much difference between the quality of the information returned by the experimental systems. Since all systems use the same retrieval architecture the results retrieved may be very similar (and for the same query identical). The techniques presented in this thesis encourage interaction with the top-ranked document set. Whilst the systems offer different interface support mechanisms (necessitated by the experimental hypotheses) they use the same underlying retrieval techniques and retrieve the same documents in response to the same queries. Friedman Rank Sum Tests were used within each subject group to test for statistically significant differences; none were significant (*inexperienced*: $\chi^2(2) = 2.34$, p = .310; *experienced*: $\chi^2(2) = 2.55$, p = .280; *overall*: $\chi^2(2) = 2.53$, p = .282). The difference between subject groups was not significant ($U(24) = 305$, p = .36) and there were no interaction effects between systems and search experience ($\chi^2(2) = .89$, p = .64) This suggests that subjects did not notice a difference in the quality of the information retrieved between systems, and this is therefore unlikely to contribute to any inter-system differences reported later in this chapter. In the next section results obtained on tasks and task categories are presented and analysed.

11.3 Tasks

As suggested in Chapter Two, the experimental search task can have a large effect on an experiment. In this section the results on the tasks attempted are presented and analysed to discern whether the tasks had an effect on subject perceptions of the experimental systems and interaction with them. Subjects were able to choose tasks from six search topics in three task categories, one task per category. In this section I analyse the reasons subjects gave for their choice, the nature of the tasks they chose and other subject perceptions. Where appropriate, I analyse the results on a per task category (i.e., pre-focus, focus-formation and post-focus) and per system basis. The results presented in this section are not directly

associated with any of the three experimental hypotheses but provide interesting insight into the experiment nonetheless.

11.3.1 Selection

The experimental design allowed subjects to choose the topic of their first search task from six options, their second topic from five options, and their third from four. [38] I was interested in *why* subjects had chosen their tasks as this may help explain anomalous findings and provide insight beneficial for the development of search tasks in future work. That is, if one can establish why subjects chose search tasks these criteria can be used to create similar tasks in the future. On the 'Search' questionnaire subjects were offered six possible explanations for their choice of task: 'interest', 'familiarity', 'no doable alternatives', 'least boring', 'no reason' and 'other'. They were asked to choose the reason that best described the rationale behind their task selection. The divided bar in Figure 11.1 illustrates the reasons given by subjects for choosing tasks.

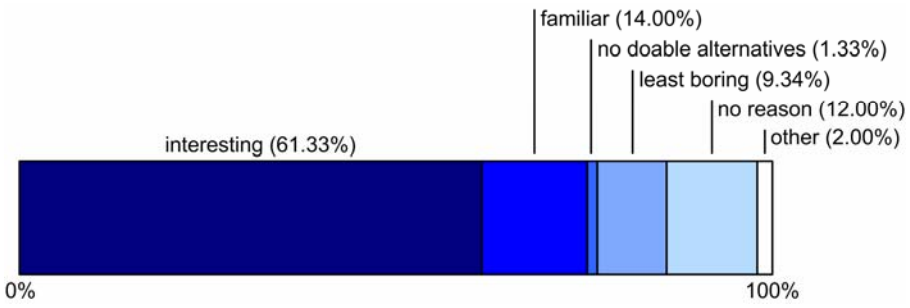


Figure 11.1. Reasons given by subjects for choosing search tasks.

The level of interest in the topic of the task appears to be the major contributory factor in deciding whether to choose a task from a number of alternatives. This supports the findings of Pilot Test 1 and the suggestion made by Borlund (2000b) that when creating tasks for interactive experimentation it is important to capture the interest of experimental subjects.

11.3.2 Nature

In this section I analyse the nature of the search tasks through subject perceptions of them generally, their perceptions of task success and the clarity of the information need created by the search tasks.

[38] Due to potential learning effects, subjects were not permitted to choose the same search topic for more than one search task.

11.3.2.1 Clarity and Complexity

Search tasks can influence subject perceptions of an experimental system or the entire experiment. For this reason it was important to determine if there were any expected or unexpected differences between tasks. Differences in the clarity and complexity of tasks between task groups were expected, since this was varied as part of the experimental design. Subjects were asked to indicate on semantic differentials how 'clear'/'unclear' and 'simple'/'complex' the tasks were. The average differential responses are shown in Table 11.3 for each task category and system type.

Table 11.3
Task characteristics across categories and experimental systems (range 1-5, lower = better).

	Inexperienced			Experienced			Overall		
Differential	Pre-focus	Focus formation	Post-focus	Pre-focus	Focus formation	Post-focus	Pre-focus	Focus formation	Post-focus
clear	3.12	2.75	**2.31**	2.96	2.80	**2.36**	3.04	2.78	**2.34**
simple	2.87	2.54	**2.01**	2.72	2.40	**1.95**	2.80	2.47	**1.98**
all (*task*)	3.00	2.65	2.16	2.84	2.60	2.16	2.92	2.63	2.16
	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
clear	1.54	1.55	**1.48**	**1.33**	**1.33**	1.37	1.44	1.44	**1.43**
simple	2.08	2.00	**1.98**	**1.92**	1.93	2.00	2.00	**1.92**	1.96
all (*system*)	1.81	1.78	1.73	1.63	1.63	1.83	1.69	1.68	1.70

n(inexperienced) = 24, n(experienced) = 24, n(overall) = 48

Pilot Test 2, described in Chapter Nine (Section 9.2.2), tested the clarity and simplicity of the tasks prior to the experiment. The pilot test showed that the tasks were all of similar levels and therefore unlikely to introduce unwanted task effects. However, it is perhaps more important to test how subjects perceived the tasks during the experiment as external factors may influence their perceptions. Table 11.3 also presents subject perceptions of the search task for different task categories and different systems. Since all tasks were created independent of the system I would expect no significant relationship between the task and system. This was verified by a Friedman Rank Sum Test applied to each differential in each subject group (all $\chi^2(2) \leq 2.41$, all $p \geq .30$).

The tasks were meant to simulate information needs at different stages in the information seeking process (ISP) and encourage different information seeking behaviours (Kuhlthau, 1991). The tasks were developed using the framework proposed by Bell and Ruthven (2004) and the complexity of the search tasks was varied as part of the experimental design. Therefore, subject perceptions of task complexity were important. The tasks were designed in such a way that the *pre-focus* task was designed to be more complex than the *focus formation*

task, which was in turn designed to be more complex than the *post-focus* task. If this was implemented successfully, I would expect a drop in the differential value for clarity and simplicity from left to right within each subject group in Table 11.3; this was generally the case. The *pre-focus* tasks were vague and required information from multiple sources. Subjects found these tasks difficult and classified tasks in this category as least 'clear' and 'simple'. The *post-focus* tasks provided subjects with more information to use to begin and conduct their search. Subjects generally found tasks in this category the more 'clear' and 'simple' than those from other categories. These findings were significant with a series of Friedman Rank Sum Tests (all $\chi^2(2) \geq 7.73$, all $p \leq .021$). Overall, the categorisation of tasks appears to concord with general subject perceptions of their clarity and simplicity. There were no significant differences between subject groups (Mann-Whitney Test, all $U(24) \leq 318$, $\alpha = .0167$, all $p \geq .24$) and no significant interaction effects between search experience and task categories (all $\chi^2(2) \leq 1.43$, all $p \geq .49$). However, there are interaction effects between search experience and systems for both differentials (*clear*: $\chi^2(2) = 1.31$, $p = .52$, *simple*: $\chi^2(2) = 1.31$, $p = .52$). The experimental systems appear to affect subject perceptions of clarity and simplicity of the search task; this affects both subject groups differently. Inexperienced subjects found searches on the Automatic system more clear and simple, perhaps because it helped them more directly. In contrast, experienced subjects found searches on the Checkbox and Recommendation systems more clear and simple, perhaps because it gave them control.

To develop a more complete picture of task effects the 'Search' questionnaire contained further questions on task success and information need clarity. I now present findings on each of these questions.

11.3.2.2 Task Success

Subject perceptions of task success are important since search systems are designed to help searchers satisfy their information needs and their desire to complete the search task they are undertaking. Also, since simulated work tasks situations were used to encourage personal relevance assessments, it is only searchers who can truly judge whether a task is complete. After each search task, subjects were asked to indicate on a five point Likert scale the extent to which they agreed with the statement *I believe I have succeeded in my performance of this task*. In Table 11.4 I present subject perceptions of task success, averaged across different groups of experimental subjects.

Table 11.4
Subject perceptions of task success (range 1-5, lower = better).

Scale	Inexperienced			Experienced			Overall		
	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
Task success	2.43	**2.23**	2.46	2.50	**2.39**	2.41	2.42	**2.31**	2.44

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

Friedman Rank Sum Tests were applied between systems on the same subject groups. For experienced subjects there were no significant inter-system differences ($\chi^2(2) = 3.67$, p = .160). However, the inter-system differences for the inexperienced subjects appeared significant ($\chi^2(2) = 8.54$, p = .014) suggesting that for this group at least one of the experimental treatments (systems) differed from the rest. The application of Dunn's *post hoc* tests revealed significant differences between the Recommendation system and the Checkbox/Automatic systems (all $Z \geq 2.01$, all p ≤ .022). Other comparisons did not reveal significant differences. The Recommendation system appears to help inexperienced subjects complete search tasks. There were no significant differences between subject groups (Mann-Whitney Test, all $U(24) \leq 322$, all p ≥ .24) and no significant interaction effects between search experience and systems ($\chi^2(2) = .70$, p = .71).

11.3.2.3 Information Need Clarity

Each subject attempted tasks from three task categories – pre-focus, focus formation and post-focus. The tasks varied in complexity, with different categories requiring information from different numbers of sources and different types of information. In Table 11.5 I present the average five point Likert scale response to the attitude statement: *I had an exact idea of the type of information I wanted*.

Table 11.5
Subject awareness of information required (range 1-5, lower = better).

Scale	Inexperienced			Experienced			Overall		
	Pre-focus	Focus formation	Post-focus	Pre-focus	Focus formation	Post-focus	Pre-focus	Focus formation	Post-focus
Awareness	2.87	2.60	**2.10**	2.54	2.12	**1.94**	2.71	2.36	**2.02**

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

As task complexity increased, subject awareness of the information required decreases. An effect of this may be that subjects are less able to choose query terms and make search decisions, and therefore need more support from the search system. Mann-Whitney Tests were applied between the independent subject groups. The results revealed significant differences for *pre-focus* ($U(24) = 399$, p = .011), *focus-formation* ($U(24) = 405$, p < .001)

and *post-focus* ($U(24) = 396$, $p = .013$) task categories. Experienced subjects appeared more aware of the type of information required during search tasks in each task category. Their enhanced search experience may mean that these subjects are better able to identify what information is necessary to complete their search.

11.3.3 Task Preference

Subjects attempted a task on each of the three systems. Afterwards they were asked to rank the tasks in their order of preference. No instructions were given on what factors to base their decision on, but subjects were asked to explain their ordering. The average subject rank for each task category is shown in the Table 11.6 for each subject group and across all subjects.

Table 11.6
Subjects' preferred task rank order (range 1-3, lower = better).

Inexperienced			Experienced			Overall		
Pre-focus	Focus formation	Post-focus	Pre-focus	Focus formation	Post-focus	Pre-focus	Focus formation	Post-focus
2.25	2.00	**1.79**	2.21	1.92	**1.92**	2.23	1.96	**1.85**

n(inexperienced) = 24, n(experienced) = 24, n(overall) = 48

A Kruskal-Wallis test was applied to the rankings in each subject group, and overall across all subjects. The results showed significant differences in the rankings assigned by inexperienced subjects ($\chi^2(2) = 11.04$, $p = .004$), experienced subjects ($\chi^2(2) = 8.85$, $p = .012$) and overall ($\chi^2(2) = 10.23$, $p = .006$). Dunn's *post hoc* tests were used to compare the task categories within each group. There were significant differences in the inexperienced group between all category pairs and in the experienced group between all pairs except *focus formation* and *post-focus* (all $Z = 1.23$, $p = .109$). Experienced subjects preferred the two less complex tasks but there was no discernable difference in the ranking between them.

I also compared the task preference between inexperienced and experienced subject groups. A Mann-Whitney Test was applied between the groups to determine the significance of any differences. The results showed that the rankings did not differ significantly ($U(24) = 347$, $p = .112$). That is, there was no discernable difference in the type of task inexperienced and experienced subjects prefer.

Subjects were asked to provide an explanation for their ranking. A variety of explanations were offered, however the most popular, in descending order of frequency were: 'interest in tasks', 'easiness of tasks', 'familiarity with similar tasks', 'task complexity', 'experimental

systems' and 'task completion'. Subjects appear to place importance on the factors that influence their ability to complete search tasks.

A deeper examination of the subject comments revealed a split between the three task categories. That is, subjects appeared to notice differences between the categories and *how* the categories differed (i.e., in complexity). Since subjects were not informed that the tasks were categorised in this way, they are making their own inferences and seem able to discern even subtle variations in task complexity. In Table 11.7 examples of the comments made by subjects are provided.

Table 11.7
Subject comments on task categories
(numbers in brackets reflect the concept/statement frequency).

Pre-focus	Focus-formation	Post-focus
1. "research-based"	1. "more focused" (2)	1. "knew what to expect"
2. "complex" (2)	2. "hard to make initial query"	2. "clear" (3)
3. "very loose"	3. "specific topic"	3. "more technical"
4. "not very specific"		4. "easy to make initial query"
5. "hard to make initial query"		5. "precise information" (2)
6. "required further interaction"		6. "know exactly what I looked for"
7. "didn't know where to look"		7. "specific topic"
8. "open subject"		8. "more effective for queries"
9. "hard to find exact information"		

$n = 48$

As can be seen from the selection of comments, subjects appeared able to determine that tasks in the three categories differed in complexity. The difference between the comments in the *pre-* and *post-focus* categories is more apparent than other pair-wise differences. Subjects were not asked specifically about the nature of the task so not all subjects provided feedback of this kind. Others chose to make reference to the information retrieved by the experimental system, their own previous search experiences and task specifics (e.g., one subject chose to write "did you know there are 18,000 dust mites in one gram of dust?").

In this section the search process and search tasks attempted by subjects have been analysed. Since these factors affect subject perceptions of the experimental systems and the experiment as a whole it is important to consider them in an analysis such as this. The search tasks play a vital role in facilitating interaction with the search systems. Therefore, it was important to establish why tasks were chosen and whether the task categories were interpreted by subjects as they were meant to be (i.e., whether the level of task complexity as perceived by subjects

matched that intended in the task categorisation). The findings presented in this section demonstrate that the Recommendation system leads to a more pleasant search and subject perceptions match the task categorisation. In what follows in this chapter I present and analyse results related to each of the three experimental hypotheses. In the next section I begin with the first, interface support.

11.4 Hypothesis 1: Interface Support

This section presents results related to the first experimental hypothesis: *the interface support provided by the experimental systems was liked by subjects and facilitated effective information access.* This hypothesis was divided into a number of sub-hypotheses that are tested in this section. To test these I analyse results obtained from a combination of questionnaire responses, system logs, informal subject comments, and my own observations. The interface support provided by all three experimental systems is compared based on how new queries are constructed, how retrieval strategies are chosen, how relevance information is conveyed and how (where appropriate) the system notified the subject of decisions it makes. The main differences between the three experimental systems are in the control they give subjects over aspects of their search.

11.4.1 Relevance Paths and Content

All systems present a large amount of information at, what I have referred to as, 'content-rich' search interfaces. Subjects were asked to express their opinion of this content in the 'Search' questionnaire and informally at the end of the experiment. As there are no path and content differences *between* systems, I only compare results between subject groups (i.e., inexperienced *versus* experienced).

From observations and informal post-search interviews, subjects appeared to use the relevance paths and found the increased levels of content shown at the search interface of value in their search. This is important, as the success of the both systems – especially the Recommendation and Automatic systems – is dependent on using these interface components. All experimental systems encouraged subjects to interact with the results of their search. They show many representations of the top-ranked documents directly to the subject at the results interface. These interfaces aim to facilitate the swift resolution of information needs but since they are novel, depend on their usability. For this reason, the training strategy (described in Chapter Nine, Section 9.10) was important, as was subject reaction to the systems. In this section results are presented on the relevance paths and information displayed at the search interface.

11.4.1.1 Relevance Paths

Subject interaction with relevance paths was automatically logged by the experimental systems. In this section I present the results of this log data analysis. Table 11.8 shows the most common path taken, the average number of steps followed, the average number of complete and partial paths and the average number of occasions where a subject went straight to a document from the first representation they visited. All averages are for each group of subjects over all search tasks. A complete path involved a subject visiting all five document representations and *then* the document itself. In partial paths, subjects visit only some document representations and do not have to visit the source document. Analysis of this sort can reveal how subjects actually used a search system rather than their perceptions of its use.

Table 11.8
Use of relevance paths (range 1-5, lower = better).

Factor	Inexperienced			Experienced		
	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
Most common path	TRS ↓ Title	Title ↓ Summary ↓ Summary Sentence		TRS ↓ Title ↓ Summary	Title ↓ Summary ↓ Summary Sentence ↓ Summary Sentence in Context	
Average steps	2.32	3.08	3.10	3.63	4.38	4.41
Average complete (partial) paths	5.20 (13.30)	7.35 (11.33)	7.54 (11.90)	7.32 (17.54)	11.64 (15.10)	11.85 (15.32)
Straight to document	6.61	6.35	6.48	9.62	9.30	9.76

n(inexperienced) = 24, *n*(experienced) = 24

Experienced subjects interacted more with the results of their search. Their paths were generally longer and they also followed more complete and partial relevance paths. They also went directly to more documents than the inexperienced subjects. These differences between groups were significant with a Mann-Whitney Test ($U(24) = 417$, $\alpha = .0167$, $p = .004$) for each pair-wise comparison (e.g., average steps (*inexperienced*/S_{Check}) *versus* average steps (*experienced*/S_{Check})). The option to directly indicate which items are relevant had an obvious effect on the interaction of experimental subjects. In the Checkbox system both subject groups interacted with shorter relevant paths than the Recommendation and Automatic systems. All users of the Checkbox system followed less complete and more partial paths than the other systems (Friedman Rank Sum Test, $\chi^2(2) = 12.43$, $p = .002$). This could be

because subjects were trying to identify which representations were relevant rather than engaging themselves fully in their search.

There were only minor differences in the use of relevance paths for different task categories. I posit that the 15 minute task time was insufficient for real differences in subject search behaviour to emerge. Those studies that have found different search behaviours for different stages in the information seeking process (e.g., Kuhlthau, 1991) have been longitudinal and have monitored search behaviours over a period of weeks and months. While subject perceptions of the tasks differed, there was insufficient evidence from their interaction to suggest they interacted differently.

11.4.1.2 Content

To test the value of the interfaces to the experimental systems, subjects were asked about how the information was presented at the results interface. A set of four semantic differentials were used to elicit subject opinion: 'helpful'/'unhelpful', 'useful'/'not useful', 'effective'/'ineffective', 'distracting'/'not distracting'. This was an important question, if subjects did not perceive direct benefit from the interfaces it may have adversely affected how they used them. The average responses for the four semantic differentials are shown in Table 11.9.

Table 11.9
Subject perceptions of information presented at the search interface
(range 1-5, lower = better).

	Inexperienced			Experienced			Overall		
Differential	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
helpful	2.07	**1.96**	2.11	2.17	**2.14**	2.17	2.17	**2.05**	2.14
useful	2.29	2.29	**2.28**	2.18	**2.12**	2.08	2.33	2.20	**2.18**
effective	2.23	2.13	**2.10**	2.34	**2.26**	2.29	2.29	**2.19**	2.20
not distracting	2.38	2.21	**2.00**	2.28	2.18	**2.17**	2.28	2.19	**2.08**
all	2.25	2.15	2.13	2.24	2.18	2.18	2.27	2.16	2.15

n(inexperienced) = 24, n(experienced) = 24, n(overall) = 48

The experimental systems presented information on the interface in the same way. Friedman Rank Sum Tests were applied within each subject group to test for statistical differences between the experimental systems and to see if components that varied between systems affected subject perceptions of the content shown. These tests revealed no significant differences in the value of the content presented between any of the experimental systems (all $\chi^2(2) \leq 2.93$, $\alpha = .0125$, all $p \geq .231$). Variations in interface provision for creating queries and making new search decisions therefore did not effect subject perceptions of how useful

the content shown to them was. There were no significant differences between subject groups (Mann-Whitney Test, all $U(24) \leq 338$, all $p \geq .15$) and no significant interaction effects between search experience and systems ($\chi^2(2) = .77$, $p = .68$). In the next section the interface techniques used to reformulate the query are evaluated.

11.4.2 Term Selection

At any point in the search the experimental systems allowed the formulation of new query statements. When prompted, the Checkbox system presented the original query and the best non-query terms in a text box and allowed the subject to retain those terms added, add their own terms or remove terms to formulate the new query. The Recommendation system presents a list of recommended terms and allows the subject to add the best terms from this list to the query. The Automatic system generates a new non-editable query, but does allow the subject to create their own query for re-searching the Web. Subjects were asked to indicate on a Likert scale how comfortable they were with each query formulation method. The average responses are shown in Table 11.10.

Table 11.10
Subject perceptions of term selection methods (range 1-5, lower = better).

Inexperienced			Experienced			Overall		
S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
2.79	**2.13**	2.96	2.63	**1.96**	2.88	2.71	**2.04**	2.92

n(inexperienced) = 24, n(experienced) = 24, n(overall) = 48

A Friedman Rank Sum Test was applied to the values in each group and the results indicated statistically significant differences in all groups (all $\chi^2(2) \geq 17.03$, all $p < .001$). Dunn's *post hoc* tests were applied to the data and revealed (in all three groups) significant differences between the Recommendation system and the other systems (all $Z = 3.12$, all $p < .001$). The differences between the Checkbox and Automatic systems were not significant in any groups (all $Z \leq 1.16$, all $p \geq .123$). In Chapter Four, the *TRSFeedback* study showed that relevance indications communicated implicitly could be a substitute for their explicit counterpart. This finding suggests in certain circumstances term selection components in such systems may also in some way be substitutable, and the case of the Recommendation system, perform better. There were no significant differences between subject groups (Mann-Whitney Test, all $U(24) = 353$, all $p = .09$) and no significant interaction effects between search experience and systems ($\chi^2(2) = 1.06$, $p = .59$).

The Likert scale analysed in Table 11.10 asks subjects to make a value judgement on the interface technique used to create the new query. Subjects appeared to like the presentation of

the terms in a list separated from the query, allowing them to choose which terms were relevant and move these terms into the query. In the Checkbox system the new terms were included in the query box meaning the subject had to remove those that were not relevant. Also, the Checkbox system required subjects to explicitly request support with query formulation, something they forgot about or appeared unwilling to do. Experimental subjects generally did not like these additional burdens. In the next section I present and analyse findings on the interface support mechanisms for retrieval strategy selection.

11.4.3 Retrieval Strategy Selection

The experimental systems implemented retrieval strategies to gather a new set of documents or restructure the information already retrieved. The Automatic system follows strategies on behalf of subjects, the Recommendation system recommends them and the Checkbox system relies on the subject to choose them. In a similar way to the previous section, subjects were asked to indicate on a Likert scale how comfortable they were with the method used to select retrieval strategies in the experimental systems. Subjects' average response for each system, from each subject group, in shown in Table 11.11.

Table 11.11
Subject perceptions of retrieval strategy selection methods (range 1-5, lower = better).

Inexperienced			Experienced			Overall		
S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
2.23	**2.04**	2.92	2.21	**1.94**	2.63	2.22	**1.99**	2.78

n(inexperienced) = 24, n(experienced) = 24, n(overall) = 48

A Friedman Rank Sum Test was applied to the values in each group and the results indicated the presence of effects in all groups (all $\chi^2(2) \geq 14.26$, all $p < .001$). Dunn's *post hoc* tests were applied to the data and revealed (in all groups) significant differences between all systems and all other systems (all $p \leq .001$). There were no significant differences between subject groups (Mann-Whitney Test, $U(24) = 350$, $p = .10$) and no significant interaction effects between search experience and systems ($\chi^2(2) = 1.94$, $p = .38$). Subjects preferred the Recommendation and Checkbox systems since they had final control over how the revised query was used. The Recommendation system was preferred since as well as giving searchers control, it also made recommendations about which strategy should be followed; subjects could ignore or accept the recommendation. Later in this chapter I use interaction logs to analyse how many of the recommended actions were accepted. The Automatic system was not liked because it removed this control and intruded on subjects' search. The option to reverse all search decisions it made did not compensate subjects for the additional burden of having to do so.

The experimental systems used different methods to gather relevance information. Some gather assessments unobtrusively from subject interaction and others more directly. In the next section I analyse the results obtained when subjects were asked about the provision of relevance information in each of the three experimental systems.

11.4.4 Relevance Assessment

The experimental systems differ in how subjects could communicate which information presented at the interface was relevant. The Checkbox system presents checkboxes next to each representation and allows subjects to explicitly mark relevant items. The Recommendation and Automatic systems use implicit assessments of relevance, generated during subject interaction with the system. Subjects were asked about how they told the system which items (e.g., titles, summaries, Top-Ranking Sentences) were relevant. Unlike traditional RF systems, subjects were not able to mark whole documents as relevant; instead they assessed representations of documents. This may allow them to make more accurate relevance assessments.

They were asked to complete two semantic differentials about:

1. the *effectiveness* of the assessment method i.e., *How you conveyed relevance to the system was*: 'easy'/'difficult', 'effective'/'ineffective', 'useful'/'not useful'.
2. how subjects *felt* about the assessment method i.e., *How you conveyed relevance to the system made you feel:* 'comfortable'/'uncomfortable', 'in control'/'not in control'.

The average obtained differential values are shown in Table 11.12 for inexperienced subjects, experienced subjects and all subjects, regardless of search experience. The value corresponding to the differential 'all' represents the mean of differentials one and two for a particular experimental system.

Table 11.12
Subject perceptions of relevance assessment methods (range 1-5, lower = better).

	Inexperienced			Experienced			Overall		
Differential	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
easy	2.46	1.88	**1.79**	2.46	2.00	**1.96**	2.46	1.94	**1.88**
effective	2.75	**1.96**	2.67	2.63	**2.18**	2.67	2.69	**2.07**	2.67
useful	2.50	**2.13**	2.42	2.46	**2.14**	2.40	2.48	**2.12**	2.41
all (*diff. 1*)	2.57	1.99	2.29	2.52	2.11	2.34	2.55	2.05	2.32
comfortable	2.46	**1.88**	2.21	**2.14**	2.21	2.26	2.30	**2.05**	2.23
in control	**1.96**	2.25	3.21	**1.98**	2.13	3.14	**1.97**	2.19	3.13
all (*diff. 2*)	2.21	2.06	2.71	2.06	2.17	2.70	2.13	2.12	2.68

n(inexperienced) = 24, n(experienced) = 24, n(overall) = 48

Friedman Rank Sum Tests were applied within each subject group (differential 1: $\alpha = .0167$, differential 2: $\alpha = .0250$). The results of this analysis suggested significant differences in all semantic differentials and all subject groups (all $\chi^2(2) \geq 10.60$, all $p \leq .005$) except the 'comfortable'/*experienced* comparisons ($\chi^2(2) = 4.21$, $p = .122$). Experienced subjects appear equally comfortable with the relevance assessments in all systems. [39] Their search experience may allow them to adapt between interface technologies more easily. Dunn's *post hoc* tests were run on all differentials revealing significant differences for all comparisons (all $Z \geq 2.26$, all $p \leq .012$). These differences suggest that subjects found the implicit methods easy and useful in their search. In the Checkbox system subjects could decide which document representations were marked as relevant. Subjects felt more in control when given the additional responsibility for communicating relevance but, for inexperienced subjects, not necessarily more comfortable. Inexperienced subjects found the explicit communication of relevance difficult. Subjects with less search experience may find it problematic to adapt to new techniques for controlling their search.

The Recommendation and Automatic systems used implicit feedback techniques to estimate which information was relevant. These systems made inferences about information needs directly from search behaviour. The systems assume that when searching for information a user will try to maximise their rate of gain of relevant information. This assumption is at the centre of *information foraging theory* (Pirolli and Card, 1995), and assumes: (i) that the examination of documents and related information is driven by information needs, and; (ii) that searchers will try to maximise their rate of gain of relevant information whilst minimising the amount of irrelevant information. To test whether information needs drove interaction in the experimental systems, subjects were asked to indicate on a Likert scale the extent to which they agreed with the statement: *As I searched, I tried to only view information related to the search task*. The average Likert scale responses are presented in Table 11.13.

Table 11.13
Subjects tried to view relevant information (range 1-5, lower = better).

Inexperienced			Experienced			Overall		
S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
1.71	**1.67**	1.78	1.71	**1.50**	1.62	1.71	**1.59**	1.70

n(inexperienced) = 24, n(experienced) = 24, n(overall) = 48

For the Recommendation and Automatic systems, these findings were important since they operate under the assumption that subjects will look try to view relevant information as they

[39] There was an interaction effect between search experience and the experimental systems for the 'comfortable' differential ($\chi^2(2) = 7.38$, $p = .025$).

search. Friedman Rank Sum Tests were applied and suggested no significant differences between systems for inexperienced subjects ($\chi^2(2) = 2.69$, $p = .261$) but there were for experienced subjects ($\chi^2(2) = 6.95$, $p = .031$). Dunn's *post hoc* tests revealed differences between the systems that gathered relevance information implicitly and the Checkbox system. Experienced subjects may have been able to infer how the Recommendation and Automatic systems choose additional terms (i.e., through the document representations viewed). There were no significant differences between subject groups (Mann-Whitney Test, $U(24) = 356$, $p = .08$) and no significant interaction effects between search experience and systems ($\chi^2(2) = .58$, $p = .75$). In the post-experiment 'Exit' questionnaire a number of experienced subjects explained that they had tried to be selective with the information they viewed since they assumed this must be how the systems that use implicit feedback gathered their evidence. That is, experimental subjects' perceptions of system operation influenced their interaction.

To assume that all the information a subject expresses an interest is in relevant may be too coarse grained since subjects can also interact with non-relevant information. To investigate the validity of this claim, interaction log data was used to calculate the proportion of all possible representations in the top 30 retrieved documents used to construct representations that were relevant (i.e., the search precision). In the Checkbox system this is the proportion of all possible representations that were marked relevant by the subject. Precision is computed in the Recommendation and Automatic systems based on the proportion of all possible representations that the subject expresses an interest in. The average number of document representations created or extracted from the top 30 documents was 320.65. There are a maximum of 14 representations per document; four Top-Ranking Sentences, one title, one summary, four summary sentences and four summary sentences in document context. However, since representations are created based on document content there is a chance that the documents may contain insufficient text to extract four sentences or may take too long to download. The precision values are shown in Table 11.14 and in Figure 11.3. For the Checkbox system the potential precision value is also given (in brackets) if implicit assessments had been used.

Table 11.14
Average search precision (values are percentages).

Inexperienced			Experienced			Overall		
S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
1.25 (20.96)	21.65	21.36	2.76 (17.05)	17.17	16.52	2.01 (19.01)	19.41	18.94

n(inexperienced) = 24, n(experienced) = 24, n(overall) = 48

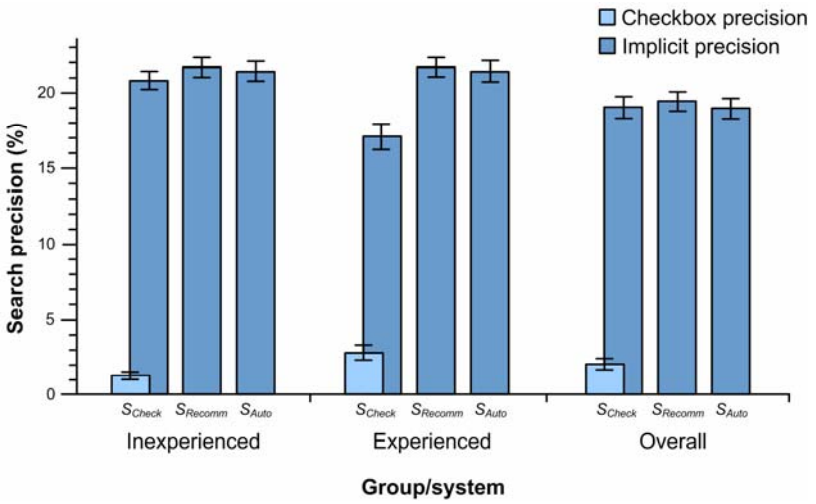


Figure 11.2. Search precision across system type and subject group (+/– *SE*).

The average search precision values shown in Table 11.14 suggest large differences in the number of items marked relevant in the Checkbox system and those inferred relevant in the Recommendation or Automatic systems. Subject criteria for marking a representation was generally very strict. During the experiment subjects suggested that an item had to be definitely relevant before they marked it. The Checkbox precision values differ significantly from those of the Recommendation and Automatic systems for both subject groups and overall (Wilcoxon Signed-Rank Test, all $T(24) \geq 229$, all $p \leq .012$). The precision values for the Recommendation and Automatic are very similar and do not differ significantly between subject groups (Mann-Whitney Test, $U(24) = 351$, $p = .097$). From these results it is obvious that experienced subjects check more items yet look at fewer. This could be because they are interacting more efficiently or assessing the relevance of items more carefully.

The highest precision value in Table 11.14 is still less than one quarter of the possible representations in the top-ranked document set. The probabilistic framework tries to estimate subject interests based on terms extracted from these representations. The experienced subjects expressed an interest in less document representations than the inexperienced subjects. These differences were not significant with Mann-Whitney Test for both the Recommendation ($U(24) = 356$, $p = .08$) and Automatic systems ($U(24) = 365$, $p = .06$). Nonetheless, this partially supports the earlier claim that experienced subjects used the systems with implicit feedback more cautiously.

Subjects provided additional informal comments on the relevance assessment process during and after the experiments. From subject comments, three factors emerged as important when indicating which results were relevant: the *method* used to communicate, the *value* of the communication and the *criteria* used during the communication. The *method* describes how

relevance indications were elicited at the interface and subjects typically forgot to provide these indications. The *value* describes the perceived benefit of conveying indications and subjects generally felt the process was not worth their effort. Finally, the *criteria* employed during the communication were typically strict (i.e., results had to be completely relevant) and subjects rarely found results they regarded as relevant. How these factors are addressed is a challenge for developers of search systems that allow subjects to make relevance indications. Subjects preferred implicit relevance assessments over explicit assessments. This is beneficial for the searcher as they no longer have to be burdened with the responsibility of providing relevance assessments and for the term selection models, who receive more evidence from which to make their decisions.

When the Recommendation and Automatic systems chose terms and made search decisions they notified the searcher by displaying messages and changing the state of interface components (e.g., colour, rank order). In the next section subject perceptions of these notifications are presented and analysed.

11.4.5 Notification

The Recommendation and Automatic systems recommended/chose new search decisions for the subject as they searched. They notified the subject through a message at the interface and by placing an 'idea bulb' next to the mouse cursor. In the 'Search' questionnaire, issued after tasks had been attempted on these two systems, subjects were asked to complete semantic differentials eliciting their opinion about these notification methods. The differentials asked about:

1. the *communication* of search decisions i.e., *The system communicated its action in a way that was*: 'unobtrusive'/'obtrusive', 'informative'/'uninformative', 'timely'/ 'untimely'.
2. the '*idea bulb*' i.e., *The appearance of the* '*idea bulb*' *when the system chose/recommended an action was:* 'not disruptive'/'disruptive', 'useful'/'not useful'.

The average differential responses for inexperienced subjects, experienced subjects and all subjects, regardless of search experience are shown below in Table 11.15.

Table 11.15
Subject perceptions of system notification methods (range 1-5, lower = better).

Differential	Inexperienced		Experienced		Overall	
	S_{Recomm}	S_{Auto}	S_{Recomm}	S_{Auto}	S_{Recomm}	S_{Auto}
unobtrusive	**1.96**	2.42	**1.58**	1.67	**1.77**	2.04
informative	**2.21**	2.54	**1.92**	1.96	**2.06**	2.25
timely	**2.38**	2.58	**2.38**	2.88	**2.38**	2.73
all (*diff. 1*)	2.18	2.51	1.96	2.17	2.07	2.34
not disruptive	**1.71**	1.67	**1.42**	1.71	**1.56**	1.69
useful	**1.71**	2.00	**1.67**	2.00	**1.69**	2.00
all (*diff. 2*)	1.71	1.83	1.54	1.85	1.63	1.84

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

Wilcoxon Signed-Rank Tests were applied within-subject groups (differential 1: $\alpha = .0167$, differential 2: $\alpha = .0250$). The results of this analysis showed that there were significant differences between systems for all differentials (all $T(24) \geq 227$, all $p \leq .014$). These results suggest that although subjects preferred the Recommendation system's notifications, how the Automatic system communicated its decisions were also effective. There were no significant interaction effects between search experience and the experimental systems used ($\chi^2(1) = 0.18$, $p = .67$).

At the end of the experiment subjects were asked to rank the experimental systems in order of preference. In the next section I analyse subject responses.

11.4.6 System Preference

Subjects used each of the three systems and were asked to rank them in their order of preference. No instructions were given on what factors to use when making their decision, but subjects were asked to explain their ordering. In Table 11.16 I present the rank order of the systems for each subject group and within this group for the different task types.

Table 11.16
Rank order of systems (range 1-3, lower = better).

Inexperienced			Experienced			Overall		
S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
2.00	**1.45**	2.46	2.25	**1.29**	2.46	2.13	**1.42**	2.46

n(inexperienced) = 24, *n*(experienced) = 24, *n*(overall) = 48

A Kruskal-Wallis test was applied to the rankings in each subject group, and to both groups combined. The results presented in Table 11.16 showed significant differences in the

rankings assigned by inexperienced subjects ($\chi^2(2) = 4.61$, $p = .010$), experienced subjects ($\chi^2(2) = 14.03$, $p < .001$) and all subjects ($\chi^2(2) = 16.22$, $p < .001$). A Dunn's *post hoc* test was used to perform multiple comparisons within each subject group. There was a significant difference in the inexperienced group between the Automatic and Recommendation systems ($Z = 2.23$, $\alpha = .0167$, $p = .013$). For experienced subjects and across all subjects there are significant differences in the ranks assigned to the Recommendation system and the other two experimental systems (all $Z \geq 2.65$, all $p \leq .004$). The Recommendation system is the preferred search system for both subject groups and overall across both subject groups.

Table 11.17
Rank order of systems per subject group and task category (range 1-3, lower = better).

System	Inexperienced			Experienced			Overall		
	Pre-focus	Focus form.	Post-focus	Pre-focus	Focus form.	Post-focus	Pre-focus	Focus Form.	Post-focus
Checkbox	2.45	2.01	**1.54**	2.63	2.58	**1.55**	2.54	2.30	**1.55**
Recommendation	**1.05**	1.45	1.85	**1.05**	1.35	1.48	**1.05**	1.40	1.67
Automatic	**1.95**	2.66	2.76	**1.95**	2.66	2.76	**1.95**	2.66	2.76

n(inexperienced) = 8, *n*(experienced) = 8, *n*(overall) = 16

The effect of the different task categories on the ranking was also analysed. All within-group differences were significant (i.e., horizontally within group) (Friedman Rank Sum Test, $\chi^2(2) \geq 10.60$, $\alpha = .0167$, $p \leq .005$). There were no interaction effects between search experience and task categorisation (Friedman Rank Sum Test, $\chi^2(2) \geq 1.06$, $p \geq .59$). Each cell in the bottom three rows of Table 11.17 represents the average rank assigned by the subjects that attempted a task from that task category on that system. The results appear to indicate an association between task complexity and system preference, with systems that remove aspects of searcher control (i.e., Recommendation and Automatic system) being preferred for more complex search tasks and those that give searchers more control being preferred for less complex tasks (i.e., Checkbox system). However, since the number of trials in each cell is relatively small one must be conservative in any conclusions drawn from these results.

The reasons subjects gave for their rankings were also analysed. In a similar way as search tasks in Section 11.3.3, the reasons given by subjects are shown in Table 11.18.

Table 11.18

Subject comments on experimental systems
(numbers in brackets reflect the concept/statement frequency).

Checkbox	Recommendation	Automatic
1. "too much control"	1. "in control" (3)	1. "simple" (5)
2. "complex – better if user knows what they want"	2. "gives help, not over the user"	2. "too little control" (5)
3. "clunky"	3. "easy to operate...intuitive"	3. "not comfortable with results"
4. "too much hassle"	4. "non-obtrusive...no hassle"	4. "too objective"
5. "slow"	5. "good balance" (2)	5. "made user feel passive"
6. "too many choices" (4)	6. "didn't like choosing terms"	6. "a lot quicker"
7. "too many checkboxes"	7. "felt good!"	7. "least flexible system" (2)
8. "checking boxes is tiresome" (2)	8. "perfect blend"	8. "frustrating" (2)
9. "simple to use...felt in control"	9. "felt inclined to try [new words]"	9. "little indication of what system was doing" (3)
10. "a lot of effort" (2)	10. "simple to use, actions slightly unpredictable"	10. "not useful at all"
11. "concentrated on looking for information than checking boxes"	11. "powerful search options"	11. "no way of asking for a recommendation"
12. "forget to check boxes"	12. "didn't interfere"	
13. "added another dimension to search that could become frustrating"	13. "felt personal, as if it was understanding me"	
14. "a bit tedious"	14. "gain new insights and words"	

$n = 48$

Table 11.18 presents a general overview of comments provided by the experimental subjects. The Recommendation system receives mainly positive comments and the Checkbox and Automatic systems mainly negative. The Checkbox system offers too many options, increased the burden on the subject and interfered with the process of finding information. The consensus among subjects is that the Checkbox and Automatic systems do have good qualities: for the Checkbox system it is the control over which results are marked relevant, for the Automatic system it is the simplicity and control of the search. [40] However, despite these qualities subjects prefer the Recommendation system to the other systems.

In this section results have been presented and analysed for the first hypothesis. The results have shown that subjects preferred the experimental system that recommended terms and retrieval strategies. Subjects found the Checkbox system a hindrance in their search, that it presented them with too many choices and that it added an additional component to the search process that could become frustrating. The Automatic and Recommendation systems

[40] One subject remarked after a successful search on the Automatic system "maybe the system was better off being in control!".

provided a mechanism through which relevance information could be conveyed that was found to be straightforward and did not disrupt subjects' search patterns. Subjects were asked informally about the activity of creating queries in each of the three experimental systems; they preferred being able to select the terms used in the creation of their query. They did not like the Automatic system which did not let them refine their query for certain system operations. The selection of query words is an activity for which subjects want support from the system in proposing additional keywords. They suggested that this could be helpful where they may not be able to create good queries. Subjects viewed the creation of a new query as an important activity that they would rather control.

Subjects were also asked about selecting search strategies. The Automatic system removed all subject responsibility for selecting strategies. In a similar way to how they felt for query creation, subjects wished to retain control over the strategies employed, but responded well to recommendations made by the systems. For strategies that restructured retrieved information rather than recreating it, subjects were more willing to delegate control to the search system. That is, the amount of control subjects wished to retain was based on the predicted impact of the strategy.

In the next section I continue my analysis and present the results used to test the second experimental hypothesis.

11.5 Hypothesis 2: Information Need Detection

The second experimental hypothesis was that: *subjects found the terms chosen by the probabilistic implicit feedback framework valuable and worthwhile*. To test this hypothesis I analysed the *value* (can be helpful during a search) and *worth* (is correct and accurate) of the terms chosen by the framework. All experimental systems chose terms using the term selection model based on Jeffrey's rule of conditioning described in Chapter Seven. The results presented in this section therefore contribute to a test of the model rather than the experimental systems. Results are presented on a per system basis to test whether the interface support affected subject perceptions of the terms selected.

In Chapter Eight the retrieval effectiveness and the rate of 'learning' of the Jeffrey's term selection model was established with searcher simulations, independent of human subjects. The 'Search' questionnaire contained a section devoted to testing this hypothesis. Subjects were asked to answer a variety of semantic differentials, Likert scales and other question

types. These data collection methods were used to gauge the effectiveness of the term selection model from the subjects' perspective.

11.5.1 Perceptions and Actions

Subjects were asked to complete two semantic differentials on whether the terms chosen by the system were 'relevant'/'irrelevant' and 'useful'/'not useful'. The average differential values are presented in Table 11.19 grouped by subject group.

Table 11.19
Subject perceptions of terms chosen/recommended by the experimental systems (range 1-5, lower = better).

	Inexperienced			Experienced			Overall		
Differential	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
relevant	2.58	**2.25**	2.63	2.33	**2.04**	2.38	2.46	**2.15**	2.50
useful	2.88	**2.38**	2.88	2.33	**2.17**	2.29	2.61	**2.27**	2.58
all	2.73	2.32	2.78	2.33	2.10	2.33	2.53	2.21	2.54

n(inexperienced) = 24, n(experienced) = 24, n(overall) = 48

Friedman Rank Sum Tests were applied to each differential for each group. The result suggested the existence of significant differences (all $\chi^2(2) \geq 7.54$, $\alpha = .025$, all $p \leq .023$). No Dunn's *post hoc* tests revealed significant differences in all subject groups between the Recommendation system and other experimental systems (all $Z \geq 2.17$, all $p \leq .015$). This suggests that the subjects perceive the terms recommended by the Recommendation system to be more relevant and useful. Although the same term selection model is used to choose terms, the data in Table 11.8 shows that the subjects interact more with the Recommendation system, providing it with more evidence. This suggests that subjects did not notice a difference in the quality of the information retrieved between systems. The differences between subject groups were significant ($U(24) = 385$, $p = .023$) suggesting that experienced subjects responded more positively to the terms selected. There were no interaction effects between systems and search experience ($\chi^2(2) = 1.88$, $p = .39$).

To build effective query modification techniques and improve the model in future work, it is vital to not only establish which terms were relevant, but *why* they were relevant. The Checkbox and Recommendation systems offered additional terms to the subject. These terms were presented in such a way that they could be edited. I regarded the act of not removing a term (Checkbox system) and moving a term from the recommended list into the query (Recommendation system) as a sign of acceptance of that term. Subjects were asked to explain why they had accepted any of the terms recommended to them. The options available

were that: 'they meant the same', 'related to words chosen already', 'could not find better words', 'represented new ideas', 'other'. Subjects were told they could select as many options as were appropriate. In Table 11.20 the reasons given by all subjects for accepting terms recommended to them are presented.

Table 11.20
Reasons for accepting terms (values are percentages).

Reason	Inexperienced		Experienced		Overall		All
	S_{Check}	S_{Recomm}	S_{Check}	S_{Recomm}	S_{Check}	S_{Recomm}	
Meant same	**16.32**	14.29	**14.65**	12.82	**14.98**	13.58	14.27
Related words	**45.95**	40.48	**43.90**	43.59	**44.87**	41.98	43.40
No better words	**13.51**	11.90	**12.20**	10.26	**12.82**	11.11	11.95
New ideas	23.98	**30.95**	28.23	**30.77**	26.70	**30.86**	15.72
Other	0.24	**2.38**	1.02	**2.56**	0.63	**2.47**	1.26

n(inexperienced) = 24, n(experienced) = 24, n(overall) = 48

The removal of the third system meant that the analysis must be applied for a 2 × 2 factorial design. Wilcoxon Signed-Rank Tests were applied to test the significance of the data within each subject group. Most differences were not statistically significant at this level (most $T(24) \leq 185$, $\alpha = .0125$, $p \geq .16$). However, the results suggest that the Recommendation system provides more new ideas than the Checkbox system ($T(24) = 234$, $\alpha = .0125$, $p = .008$). The larger number of terms offered or other aspects of the interface support may explain these differences. There were no significant differences between subject groups ($U(24) = 319$, $p = .26$) or interaction effects between search experience and system ($\chi^2(1) = .26$, $p = .61$). From these findings, I can propose that the relatedness to current query terms and the novelty of the concepts they embody are two of the main reasons why subjects accept terms chosen by search systems on their behalf.

In all systems subjects could modify their query at any point in the search. This would involve them selecting additional query terms based on tacit knowledge and their current search experience. A good term selection model should suggest relevant terms and suggest terms that initiate ideas for other terms. In this investigation subjects were asked to describe where the additional terms *they entered* originated. They could select one from 'list of terms suggested by the system', 'retrieved set of documents and extracted information', 'a combination of the first two' and 'other'. If subjects chose 'other' they were asked to provide more details. Table 11.21 shows the origins of new terms entered by the subject. The values in the table are percentages and the sum of each column is 100%.

Table 11.21
Origin of additional terms (values are percentages).

Source	Inexperienced			Experienced			Overall			All
	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	
System terms	8.33	**20.84**	16.67	**29.17**	20.84	**29.17**	18.75	20.84	**22.92**	20.83
Documents and Extracted Information	20.84	**25.00**	16.67	29.17	**33.33**	16.67	25.00	**29.17**	16.67	23.62
Combination of the above	**50.00**	45.83	45.83	12.50	**33.33**	12.50	31.25	**39.57**	29.17	33.33
Other	**20.83**	8.33	**20.83**	29.16	12.50	**41.66**	25.00	10.42	**31.24**	22.22

n(inexperienced) = 24, n(experienced) = 24, n(overall) = 48

Most subjects appeared to choose additional terms based on a combination of the terms chosen by the system and the documents and extracted information. This is a worthwhile finding as it shows the terms generated by the model are not only useful to represent current information needs but to facilitate their development. Friedman Rank Sum Tests were conducted for each differential within each subject group. The results implied the existence of statistically significant differences in each group (all $\chi^2(2) \geq 9.92$, $\alpha = .0125$, all $p \leq .007$). The high percentage of new ideas from 'other' sources (the percentages shown in the last row of Table 11.21) came from a combination of the simulated work task situation and the subject's tacit knowledge. The differences between the subject groups is significant for all differentials (all $U(24) \geq 392$, $\alpha = .0125$, all $p \leq .016$) . There is also evidence of interaction effects between the level of search experience and the experimental systems for the 'combination of the above' and 'other' differentials ($\chi^2(2) \geq 5.80$, $\alpha = .0125$, all $p \leq .002$). This suggests that the level of search experience affects where subjects get their terms and that this source varies depending on the experimental system.

The findings show that in systems that removed subject control, subjects were more likely to use the words proposed to initiate new ideas and search directions. The Checkbox system was dependent on subjects marking results as relevant. As a consequence, the words suggested were from items the subject already knew were relevant. Systems that remove subject control over creating queries may be most appropriate for encouraging new and potentially useful search directions. This can be helpful if the subject is struggling with their search. Whilst subjects want to retain control over the additional words used, it may not be in their interests to do so.

The findings also show that the amount of interactivity in how additional terms were chosen influences where the terms were chosen from. When given less control, subjects were more

likely use the system's words or other sources such as the task, tacit knowledge or previous search experience. However, subjects did not use the documents or extracted information as inspiration for new query terms. Subjects depend on the Automatic system to reorder documents and Top-Ranking Sentences; subjects did not have any control over those activities in that system. I can conjecture that when subjects could not manipulate the space in which they searched, they were less likely to use that space to assist them in constructing new queries.

A good term selection model should select terms on behalf of the subject that approximate their information needs. To be used effectively subjects must trust the systems to select appropriate terms. Subjects were asked whether they trusted the system to choose terms on their behalf. They completed a Likert scale to indicate the extent they agreed with the statement: *I would trust the system to choose words for me*. A summary of responses is provided in Table 11.22.

Table 11.22
Trust system to choose terms (range 1-5, lower = better).

Inexperienced			Experienced			Overall		
S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}	S_{Check}	S_{Recomm}	S_{Auto}
2.19	**2.03**	2.48	2.19	**1.65**	2.19	2.19	**1.84**	2.34

n(inexperienced) = 24, n(experienced) = 24, n(overall) = 48

Friedman Rank Sum Tests were conducted for each differential within each subject group. The results suggested the existence of statistically significant pairs (all $\chi^2(2) \geq 11.24$, all $p \leq .001$). Dunn's *post hoc* tests revealed that there were significant differences in all inexperienced comparisons and for the experienced and overall subject groups, the Recommendation/Automatic (*experienced*: $Z = 2.03$, $p = .021$; *overall*: $Z = 2.00$, $p = .023$) and Recommendation/Checkbox (*experienced*: $Z = 2.05$, $p = .020$; *overall*: $Z = 1.90$, $p = .029$). Subjects appear to trust systems that give them control over query modification more than those without this facility.

Subjects were encouraged to provide comments on the terms suggested by all three systems. In general the feedback received was encouraging. Some subjects complained that certain terms and their plural appeared in the query suggested by the system (e.g., 'mite', 'mites'), this was unhelpful. On the other hand, one of the search tasks involved looking for art galleries in Rome. Since some of the retrieved pages were in Italian the system would occasionally suggest Italian words (e.g., 'galleria', 'museo') that were regarded by subjects as useful for their search. The system is therefore suggesting terms that the searcher may be

incapable of selecting. In general subjects responded well to the terms chosen or recommended by the framework. The terms selected were helpful in either reinforcing current ideas or providing new ideas from which to advance their search. In the next section the interaction logs generated by each experimental system are analysed to provide further insight into how the framework was used in this experiment.

11.5.2 Query Construction

I this section I use interaction logs generated by each system to further investigate the creation of query statements. Since each experimental system supports different term selection strategies then different log data is available for each system. In this section the results from system logs are presented. The Automatic system does not allow the user the option of directly changing the new query. For this reason the logs analysed in this section are from the Checkbox and Recommendation systems.

Both systems use the probabilistic framework (Chapter Seven) for selecting query modification terms. The Checkbox system relies on the subject to mark items as relevant then suggests new query terms when instructed. The Recommendation system uses implicit feedback and recommends a list of terms to the subject.

11.5.2.1 Checkbox system

Unlike the other experimental systems, the Checkbox system awaits instruction from the subject before offering assistance. When requested, the system chooses the best six terms and appends them to the current query. The searcher then has the option to edit the query; adding or removing terms. I regard the removal of a term from those added by the system as a sign of dissatisfaction with the term (and its retention as a sign of satisfaction). Therefore, I use the proportion of terms added/removed from the original query as an indication of satisfaction/dissatisfaction with the term selection component of the probabilistic implicit feedback framework. Across all tasks on the Checkbox system an average of 2.15 of the six terms (35.83%) were rejected and 3.85 (64.17%) of terms were retained.

11.5.2.2 Recommendation system

The Recommendation system presents a list of recommended terms and allows subjects to choose terms from this list and add them to their query. This list contains 20 terms and it is unreasonable to expect subjects to add all 20 terms to their query. [41] It is also unreasonable to

[41] Due to limits with the underlying search system, a query used to re-search the Web cannot be any longer than 10 terms.

measure the proportion of the list that is selected as this is not comparable with the results in the previous section. Instead, I consider the quartiles of the list, and *where* in the list the terms reside. In a similar way to Efthimiadis (1993) it is possible to measure the proportion of offered terms added from the four quartiles of the list (top, top-middle, bottom-middle, bottom). Figure 11.4 shows how the top 20 terms were divided into four quartiles. The part of the list in the figure with a scrollbar represents the six terms shown to the searcher at any particular time. Subjects were not told that the terms in the list were ranked in descending order meaning they may not expect higher ranked terms to be more relevant. This allowed a more robust analysis of the list ordering, as if subjects chose more terms from near the top it would be because they thought they were useful, not that they assumed they should be.

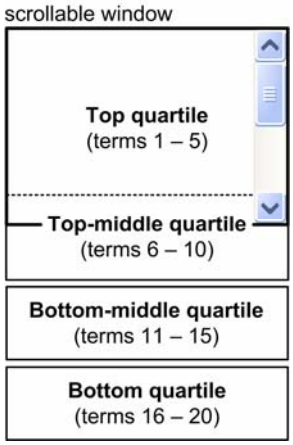


Figure 11.4. Four quartiles of the Recommendation system term list.

Since the terms are ranked by the framework, the location of terms in the list can give a clue about how well the term selection model operates. In Table 11.23 the proportion of terms chosen from each quartile in the list is shown for different subject groups and overall across all subject groups. The values in the table are percentages of the whole list.

Table 11.23
Proportion of terms chosen from list quartiles (Recommendation system only).

Quartile	Inexperienced	Experienced	Overall
Top	54.75	47.05	51.40
Top-middle	25.00	24.78	24.89
Bottom-middle	10.25	19.22	14.24
Bottom	10.00	8.95	9.47

n(inexperienced) = 24, n(experienced) = 24, n(overall) = 48

Mann-Whitney Tests were conducted for each quartile between each subject group (α = .0125). The results were significant for the top ($U(24) = 401$, p = .001) and bottom-middle

($U(24) = 401$, $p = .001$) quartiles, but not for the top-middle ($U(24) = 321$, $p = .248$) or bottom ($U(24) = 341$, $p = .137$). Overall subjects chose more than half of the recommended terms from the top five and over three-quarters (75.29%) from the top 10 terms (i.e., top and top-middle quartiles collectively). This implies that the subjects generally agreed with the ranking of terms by the term selection model.

To allow for the differences between the number of terms presented and more fully evaluate the recommended list of terms I ignore the scrollbar and only analyse terms with an initial rank position in the first six i.e., only terms that initially appear in the recommended list without the need to scroll. This meant that term selection methods in the Checkbox and Recommendation systems could be compared. Across all tasks and subject groups an average of 3.95 terms from the top six terms (65.85%) were added to the query. Analysis of these findings showed that although there was no significant difference between the number of terms added in the Recommendation and Checkbox systems (with a Wilcoxon Signed-Rank Test, $T(24) = 198$, $p = .087$). The way that additional terms are offered to subjects at the interface appears to only have a slight effect on the number of terms accepted.

11.5.2.3 Automatic system

Other than re-searching the Web, there was no mechanism for direct query refinement in the Automatic system. Subjects could modify and submit a new query to the system (i.e., re-search the Web), but received no support in choosing the terms to comprise this query. The queries submitted by subjects for the re-searching operation were typically smaller in this system (where the subject received no support) than in the other experimental systems which offered subjects assistance (2.53 terms *versus* 5.43 terms). The systems that implemented mechanisms for interactive query modification allowed subjects to build richer queries for generating new sets of search results.

In this section I have presented and analysed findings to test the second experimental hypothesis. The results have shown that the term selection model in the probabilistic framework chooses terms that are relevant and useful to subjects. The results also show that the nature of the interface support can affect subject perceptions of model effectiveness, including how much trust they place in it to choose terms on their behalf. In the next section I present and analyse results on the component used to estimate information need change that is used in the Recommendation and Automatic systems.

11.6 Hypothesis 3: Information Need Tracking

This section presents results related to the third experimental hypothesis: *subjects found the retrieval strategies chosen by the probabilistic implicit feedback framework valuable and worthwhile*. In the Recommendation and Automatic systems a component works in the background to suggest or choose new retrieval strategies during the search. These strategies are selected based on the extent of changes in the search system's formulation of information needs (i.e., changes in the list of candidate terms from which the system chooses query modification terms). To do this, the system uses the information need tracking component from the probabilistic implicit feedback framework described in Chapter Seven. As suggested in earlier chapters the framework can either re-search the Web or reorganise the information already retrieved. I test the effectiveness of this component using Likert scale and semantic differential responses, system logging (e.g., the proportion of system search decisions that are accepted by the subject) and informal subject comments.

11.6.1 Perceptions and Actions

In the 'Search' questionnaire, completed after each search task, subjects were asked to indicate on a five point Likert scale how often the retrieval strategy chosen by the framework reflected the changes in the information they were searching for. In the training session it was made clear to subjects that this change did not have to be a change in topic, it could simply be a refinement of their current search. They were asked to provide an assessment on a scale between 'never' and 'always'. The average scale responses are shown in Table 11.24.

Table 11.24
Subject perceptions on the appropriateness of retrieval strategy (range 1-5, lower = better).

Inexperienced		Experienced		All	
S_{Recomm}	S_{Auto}	S_{Recomm}	S_{Auto}	S_{Recomm}	S_{Auto}
2.54	2.58	**2.67**	2.71	**2.60**	2.65

n(inexperienced) = 24, n(experienced) = 24, n(overall) = 48

The within and between group differences were not significant (*within*: Wilcoxon Signed-Rank Tests, all $T(24) \leq 182$, all $p \geq .180$; *between*: Mann-Whitney Tests, all $U(24) \leq 358$, all $p \geq .08$) and there were no interaction effects between search experience and experimental systems ($\chi^2(1) = 0.26$, $p = .61$). Since the mechanism for selecting retrieval strategies was the same between systems it was expected that that subject perceptions of the strategies would be similar. This was the case, but subjects again appeared slightly more positive about systems that gave them ultimate control over interface decisions.

To further test the information need tracking component, subjects were asked about the retrieval strategy chosen or recommended by the experimental system. A set of three semantic differentials were used to elicit subject opinion: 'useful'/'not useful', 'helpful'/'unhelpful', 'appropriate'/'inappropriate'. The strategy chosen by the system reflects changes in the system's estimation of the information need. The responses for the three differentials are shown in Table 11.25.

Table 11.25
Subject perceptions of retrieval strategies (range 1-5, lower = better).

Differential	Inexperienced		Experienced		Overall	
	S_{Recomm}	S_{Auto}	S_{Recomm}	S_{Auto}	S_{Recomm}	S_{Auto}
useful	**2.38**	2.79	2.25	**2.21**	**2.31**	2.50
helpful	**2.54**	2.75	2.42	**2.21**	**2.48**	**2.48**
appropriate	**2.50**	2.92	**2.25**	**2.25**	**2.38**	2.58
all	2.47	2.82	2.31	2.22	2.39	2.52

n(inexperienced) = 24, n(experienced) = 24, n(overall) = 48

The 'useful' and 'helpful' differentials in Table 11.25 measure the value of the strategy, i.e., how can the strategy assist subjects to search more effectively, and the 'appropriate' differential measures its worth, i.e., how well it performs. Wilcoxon Signed-Rank Tests were applied for each differential between systems. The tests revealed significant differences within the inexperienced subject group ($T(24) = 246$, $\alpha = .0167$, $p = .003$**)** but not the experienced group ($T(24) = 209$, $\alpha = .0167$, $p = .047$**)**. Inexperienced subjects found the retrieval strategy chosen by the Recommendation system significantly more 'useful' ($Z = 2.58$, $p = .005$), 'helpful' ($Z = 2.26$, $p = .012$) and 'appropriate' ($Z = 2.41$, $p = .008$) than the Automatic system. This was anomalous since the systems used the same underlying mechanisms to choose retrieval strategies. The only difference between the systems was in how the strategy was communicated. For inexperienced subjects, the method used to communicate the decision influenced subject perceptions about the value of the strategy. Experienced subjects seem more able to isolate the mechanism behind the strategy selection and no significant differences between the differentials were discovered for that group (all $Z \leq .74$, all $p \geq .23$). That is, experienced subjects were more able to analyse the value and worth of the information need tracking component independent of the way the decisions it made were communicated.

A good information need tracking component should choose retrieval strategies that approximate changes in the information needs of searchers and assist them in finding relevant information. To be used effectively, searchers must trust the systems to select appropriate

retrieval strategies. Subjects were asked whether they trusted the system to choose retrieval strategies on their behalf. They completed a Likert scale to indicate the extent they agreed with the statement: *I would trust the system to choose an action* [42] *for me*. A summary of responses is provided in Table 11.26.

Table 11.26
Trust system to choose retrieval strategy (range 1-5, lower = better).

Inexperienced		Experienced		Overall	
S_{Recomm}	S_{Auto}	S_{Recomm}	S_{Auto}	S_{Recomm}	S_{Auto}
2.67	2.92	**2.67**	2.67	**2.67**	2.79

n(inexperienced) = 24, n(experienced) = 24, n(overall) = 48

Wilcoxon Signed-Rank Tests were applied within each subject group to compare systems and all subjects and systems compared to the mid-value of the Likert scale (i.e., 3). The results showed no significant within-group differences (all $T(24) \leq 160$, all $p \geq .390$), significant differences from the mid-value ($T(24) = 229$, $p = .012$) and no interaction effects between search experience and experimental systems ($\chi^2(1) = 0.15$, $p = .70$). Subjects reacted positively to the search strategies proposed by the system. Inexperienced subjects appeared to trust systems that gave them control over how the new query was used; for experienced subjects there was no difference.

In this section I have presented an analysis of subject perceptions of the retrieval strategy selection component. In the next section, I use system log data to analyse how subjects actually selected retrieval strategies. These logs, created as subjects searched, provide evidence to allow a deeper analysis of subject search activities.

11.6.2 Retrieval Strategy Selection

The Recommendation and Automatic systems make search decisions on subjects' behalf, whereas the Checkbox system relies on subjects to make their own decisions. Subjects are given the option to reverse the search decisions the systems made. In Table 11.27 I give the proportion of each type of action that was reversed. This reversal is regarded as an indication of dissatisfaction with the outcome of followed strategy.

[42] The word 'action' is used in the questionnaires rather than 'retrieval strategy' or 'search decision'. It was felt that subjects could relate better to 'action'.

Table 11.27
Proportion of retrieval strategies accepted or reversed (values are percentages).

Subject Action	Inexperienced		Experienced		Overall	
	S_{Recomm}	S_{Auto}	S_{Recomm}	S_{Auto}	S_{Recomm}	S_{Auto}
Accepted	72.43	**75.60**	64.67	**69.10**	68.55	**72.35**
Reversed	**27.57**	24.40	**35.33**	30.90	**31.45**	27.65

n(inexperienced) = 24, n(experienced) = 24, n(overall) = 48

The differences between the systems within the subject groups are not significant (Wilcoxon Signed-Rank Test, all $T(24) \leq 156$, all $p \geq .431$) but it is between groups (Mann-Whitney Test, all $U(24) = 399$, all $p \leq .011$). Experienced subjects tended to accept a lower number of retrieval strategies chosen by the system than inexperienced subjects. These subjects may be more reticent about search systems making decisions of this nature on their behalf and feel able to make such decisions on their own.

I use a measure known as *strategy overlap* to determine how closely the decisions made by the information need tracking component concord with subject decisions. I measure the degree of strategy overlap using the Checkbox system and the Recommendation system. The methods used in each system are slightly different. In the Checkbox system the strategy selection component runs in the background, completely invisible to the subject and not involved directly in any strategy selection decisions. That is, whilst the component chooses retrieval strategies based on changes in its formulation of information needs, these strategies are never shown to the subject and never executed. At any point in time, the component holds that retrieval strategy that it regards as most appropriate. I measure the degree of strategy overlap based on how frequently subjects choose the same strategy as the system would choose. In the Recommendation system the overlap is a measure of how many strategies followed by the subject that were also the system's recommendation at that time. This is different from the results reported in Table 11.27, since for this analysis I do not consider whether the strategy was eventually reversed or accepted. This is given as a percentage and is presented in Table 11.28 for inexperienced subjects, experienced subjects and across all subject groups.

Table 11.28
Proportion retrieval strategy overlap between system and subject (values are percentages).

System	Inexperienced	Experienced	Overall
Checkbox	61.60	57.85	59.73
Recommendation	74.66	59.32	66.99

n(inexperienced) = 24, n(experienced) = 24, n(overall) = 48

On approximately 60% of occasions the framework implemented in the Checkbox system predicted the strategy executed by the subject. The differences are not significant between subject groups with a Mann-Whitney Test ($U(24) = 345$, $p = .120$). This is a reasonable result since the evidence gathered to predict the changes that result in the strategy are based on a small amount of evidence explicitly provided by the subject through their interaction. The strategy overlap for the Recommendation system is higher than the Checkbox system. There are at least two reasons for this: (i) since it gathers relevance assessments implicitly the system has more relevance information from which to make its decisions, and (ii) the presentation of the recommendation at the interface may have unduly influenced subjects into selecting it. The inexperienced subjects follow significantly more of the system's recommendations than the experienced subjects (Mann-Whitney Test, $U(24) = 417$, $p = .004$). They may require the additional support or be less cautious than the experienced subjects about accepting it. The Checkbox system may give an artificially low strategy overlap (because of the small amount of evidence) and the Recommendation system an artificially high value (because of the influence of presenting its decisions). Therefore, I conjecture that a 'true' strategy overlap value may well lie somewhere between these two extremes.

In this section the information need tracking component of the probabilistic implicit feedback framework has been tested. Subjects were asked to comment informally about the retrieval strategies. In a similar way to how they felt for query creation subjects wished to retain control over the strategies employed, but responded well to recommendations made by the system. For strategies that restructured retrieved information rather than recreating it, subjects were more willing to delegate control to the search system. That is, the amount of control subjects wished to retain was based on the predicted impact of the strategy. Subjects suggested that the component should be more sensitive to larger changes in information needs and that it reordered documents when their intuition would have been to re-search. Nonetheless, the component performed well and the results have demonstrated that the component makes search decisions that are appropriate and that subjects find useful.

11.7 Chapter Summary

In this chapter I have presented and analysed the findings of the user experiment. The experiment aimed to compare the effectiveness of three search interfaces that varied searcher control and responsibility over aspects of the search, and test the probabilistic implicit feedback framework presented in Chapter Seven. In Table 11.29 I summarise the results for each of the sub-hypotheses described in Chapter Nine.

Table 11.29
Evidence to support experimental hypotheses.

Hypothesis	Supported?	Evidence
Hypothesis 1. Interface Support		
Relevance Paths and Content (Hypothesis 1.1) Subjects find the information presented at the interface useful.	✓	Section 11.4.1
Term selection (Hypothesis 1.2) Subjects want control in formulating new queries.	✓	Section 11.4.2
Retrieval strategy selection (Hypothesis 1.3) Subjects want control in making search decisions.	✓	Section 11.4.3
Relevance assessment (Hypothesis 1.4) Subjects want the experimental system to infer relevance from their interaction.	✓	Section 11.4.4
Notification (Hypothesis 1.5) Subjects find system notifications helpful and unobtrusive.	✓	Section 11.4.5
Hypothesis 2. Information Need Detection		
Value (Hypothesis 2.1) Query modification terms chosen by the framework are relevant and useful.	✓	Section 11.5.1
Worth (Hypothesis 2.2) Query modification terms chosen by the framework approximate subject information needs.	✓	Section 11.5.1 Section 11.5.2
Hypothesis 3. Information Need Tracking		
Value (Hypothesis 3.1) The retrieval strategies chosen by the framework are beneficial.	✓	Section 11.6.1
Worth (Hypothesis 3.2) The retrieval strategies chosen by the framework approximate changes in the information needs of subjects.	✓	Section 11.6.2

The results have shown that subjects did not like having to mark items as relevant (as in the Checkbox system) or devolving control over query creation and retrieval strategy selection (as in the Automatic system). Subjects preferred to communicate relevance implicitly, and receive system support in creating queries and making new search decisions, but still retain ultimate control over these two activities. The Recommendation system offered them the facilities to do this. Hypothesis 1 was supported by these findings

In this chapter I also evaluated the probabilistic implicit feedback framework presented in Chapter Seven, to modify queries and select retrieval strategies. Subjects found the terms and strategies selected by the framework useful, relevant and appropriate in the context of their search. Hypotheses 2 and 3 were supported by these findings. In the next chapter I discuss the implications of the results obtained.

Chapter 12

Discussion

12.1 Introduction

In the previous chapter I presented and analysed the results of the user experiment. In this chapter these results are discussed in the context of this thesis and related literature; where appropriate, the findings are also compared to those of Pilot Test 1, described in Chapter Nine. In particular, I concentrate on results that relate to the three experimental hypotheses and other parts of this thesis. Each hypothesis is addressed in turn and this chapter concludes with a summary discussion of the implications of my findings.

Selecting worthwhile terms on behalf of searchers relies on an ability to predict their information needs to a very fine level of granularity. Traditional implicit and explicit relevance feedback approaches use sets of documents from which to extract terms for query modification (Salton and Buckley, 1990; Kelly and Teevan, 2003). This approach is coarse-grained since documents can contain a large number of erroneous terms (Allan, 1995). The approaches described in this thesis utilise interaction with novel content-rich search interfaces to modify the query statements and make search decisions.

Users of traditional search systems are typically responsible for all aspects of their interaction, from the selection of query terms to the assessment of the results obtained. This can be problematic as searchers typically receive no training in how to create queries, exhibit limited interaction with the results of their searches and do not examine results closely (Jansen *et al.*, 2000). The search interfaces presented in Parts II and IV use query-relevant document representations to facilitate access to potentially useful information and encourage searchers to closely examine search results. The findings in Part II showed that increased searcher interaction with retrieved information led to more effective searching. The interfaces in Part

II use the content of the most relevant documents in the retrieved set in an approach I call content-driven information seeking (CDIS).

IR systems that use implicit feedback make inferences about what information is relevant based on searcher interaction. They do not intrude on the searcher's primary line of activity (i.e., satisfying their information need). That is, the treatment by the system of the searcher's action as evidence of relevance is secondary to the main task, which is to respond to the searcher's instruction (Furnas, 2002).

RF systems typically have functionality for choosing query words, providing relevance information and making new search decisions. In this experiment I developed three experimental systems that tested these functions with subjects with different skill levels and search experience. This chapter begins with an initial discussion of the search process and search tasks attempted by subjects, then discusses interface support and the performance of the framework in detecting and tracking information needs.

12.2 Tasks and the Search Process

In this thesis I have described a number of user studies. Most of these studies have used simulated work task situations to facilitate interaction with the experimental systems. [43] These allow subjects to make personal assessments of what constitutes relevant information and allow search systems to be compared on the same underlying information need. In the studies described in Part II the subjects were not given a choice of tasks. This led to slight problems as some subjects were not interested in the task assigned to them. Borlund (2000b) recommended that in the construction of search tasks, experimenters should consider the involvement of subjects, the application of dynamic and individual information needs (real and simulated) and the use of multidimensional and dynamic relevance judgements. Subjects with an interest in the subject area of the task are more likely to become involved in the task and form an individual perspective of it. In Pilot Test 1 and in the experiment described thus far in Part IV I offered subjects a choice of search tasks that gave subjects more control over the tasks they attempted.

In Chapter Four I discussed the use of the top-ranking sentence based experimental interfaces in relation to the model of the Information Search Process (ISP) proposed by Kuhlthau (1991). This model assumes that there is a point of 'focus' (Kelly, 1963; Belkin, 1980; Kuhlthau, 1991) where uncertainty drops and searchers can better identify the topic of their

[43] With the exception of the *TRSFeedback* study in Chapter Four.

search. The findings from the user studies described in that chapter suggest that the systems support two of the six stages of the ISP: *exploration* (investigating information on general topic) and *collection* (gathering relevant or focused information). Those systems that used implicit feedback to reorder the Top-Ranking Sentences also displayed limited support for the *formulation* stage (formulating the search focus). However, since it is the system that is refining its formulation of the information need internally, the extent to which these systems support formulation (from the searcher's perspective) is limited. Through encouraging more interactivity in query creation, the systems presented in the experiment I have described in Part IV help searchers refine their query and improve support for the formulation stage of the ISP.

In this experiment, tasks were divided into three categories based on the actions common to each stage in the ISP. The tasks used in this evaluation simulate stages before the focus, as the focus is forming and after the focus has formed. The three task types created were assigned the names: *pre-focus*, *focus formation* and *post-focus*. The pre-focus tasks encouraged subjects to locate background information, the focus formation broadly relevant information and the post-focus broadly relevant or pertinent (focused) information.

Search tasks were created for each category using the approach described by Bell and Ruthven (2004) i.e., the task categories were varied in terms of complexity. The pre-focus task was assumed to simulate the state of an information need in the initial explorative stages of a search; encouraging browsing behaviour; this task was assumed to be highly complex. The focus-formation task simulated information needs as subjects began to understand what they were looking for and could then make decisions about what information was relevant; this task was assumed to be of moderate complexity. Finally, the post-focus task simulated a well-formed information need and encouraged focused information seeking; this task was assumed to have a low complexity. It is in the pre-focus stage where the information needs are least well-defined and most changeable.

I selected six search topics to approximate real information seeking scenarios. Subjects chose a task from each category without topic repetition to limit learning effects. This meant they choose the first task from six topics, the second from five topics (the first topic could not be attempted again) and the third from four (the first and second topics could not be attempted again). This methodology meant that the third topic selected could be a subject's third preference. The effect of this was negligible and was preferred to situations where topics

were not removed (serious learning effects) [44] or subjects constructed their own search tasks (no comparability between systems). Unlike naturalistic studies (Beaulieu, 1997; Kelly, 2004) this investigation did not study natural search behaviours in operational settings. This experiment was a comparative evaluation and a deviation from a methodology where I could control many external factors could invalidate the experimental findings.

Subject comments in the 'Exit' questionnaire led me to conclude that they were able to identify differences between task categories. Subjects remarked that their search behaviour and task performance were affected by the nature of the task they attempted. It emerged from these comments that not only did subjects know that the task categories were different but that they also knew *how* the categories differed (i.e., in their complexity). There were no discernable differences in subject perceptions of the tasks between systems although subjects did find the pre-focus tasks more 'complex', 'unfamiliar' and 'unclear' than tasks from the other categories. Although there were only minor differences in subject interaction for each type of task, subject perceptions suggest that systems that gather relevance information using implicit feedback and make or recommend decisions (i.e., the Automatic or Recommendation systems) are most useful during the uncertain, formative stages of the information seeking process. Since these systems removed the burden of directly communicating relevance, subjects could focus on viewing and interpreting the documents and extracted information presented at the search interface. The Checkbox system was preferred in circumstances where subjects were more certain about the information they were searching for. When they become more aware of their information need, they felt more able to identify what information is relevant. In such situations they also *want* more control over system decisions and seem less reluctant to provide relevance feedback directly. This is a potentially significant finding, although since it does not form one of the hypotheses tested in this thesis a more complete analysis of the results are reserved for future work.

In a related study, Fowkes and Beaulieu (2000) suggested that the complexity of the search may be an indicator of when to use different query modification techniques. They found that for searches where the desired information is clearly defined and for which the searcher can retrieve relevant information they do not require as much control over the terms that comprise the query. Searches involving vague information needs or in cases where little relevant information is being retrieved benefit more from increased control over the query terms. However, in this experiment I demonstrate that the same may not be true for relevance assessments; subjects felt most comfortable directly communicating relevance to the

[44] With subjects being able to choose the same topic for all three tasks.

Checkbox system for less complex tasks. Since the direct communication of relevance information is dependent on an ability to identify what information is relevant, searchers may feel more comfortable doing this for less complex tasks. For more complex searches they may be unable to identify what information is relevant and may therefore rather rely on inferences made by the search system.

Subjects were asked about their perceptions of task success after they had attempted each search task. Task success was assessed from the subjects' perspective since this most closely reflects real life retrieval situations. Personal assessments on task completeness also fit better with the use of simulated work task situations, which require subjective judgements on what information is relevant. The findings of the experiment suggest that inexperienced subjects perceived higher levels of task success on the Recommendation system than any of the other two experimental systems. They commented that the way the system communicated its decisions (i.e., unobtrusively) meant they were not impeded in their search by a need to control the system. Subjects spent time on the Checkbox system assessing document representations for relevance, rather than searching for information and reversing or examining the effects of the Automatic system's decision. This reduced the amount of information they could examine during a search and for the more complex tasks lessened the likelihood of a successful search.

The results discussed in this section show that subjects noticed the differences in task complexity and that the experimental systems that were most proactive in offering searcher assistance were most useful for search tasks that encouraged explorative information seeking behaviour. The interface support mechanisms were the only differences between the experimental systems. In the next section I discuss findings about the first of the three research questions, which addresses the interface support issue.

12.3 Interface Support

In this section I discuss aspects of the interface support offered by the experimental systems. The three systems provided different mechanisms that varied how much control searchers had over aspects of their search. Many of the interface design decisions made for the experimental systems described in Chapter Ten arose from subject comments during Pilot Test 1. In that study two prototype experimental systems were tested: a manual baseline system and an experimental system that used implicit feedback. The manual baseline gave subjects control over the terms selected and retrieval strategy followed and the implicit feedback system automatically modified the query and used the new query to perform a new

retrieval strategy. The implicit feedback system gave subjects no control over the process other than the option to reverse system decisions. This pilot test demonstrated that the heuristic-based implicit feedback framework (from Chapter Six) could approximate searcher interests and estimate changes in these interests (i.e., the framework worked well). However, subjects also suggested that they preferred systems that offered assistance in making search decisions, but gave them final control over the choices made. These comments were considered and influenced the development of the systems used in this experiment.

Systems with three different types of interface support were used to communicate the decisions on which terms and retrieval strategies were chosen by the underlying probabilistic framework. The way in which these decisions were communicated and the level of searcher control over them, was varied between systems. In the Checkbox and Recommendation system, new query terms were suggested as a recommendation and could be edited by the subject. In contrast, the Automatic system chose terms for the subject. Subjects generally preferred the interface support mechanisms provided by the Recommendation system.

In a related study, Beaulieu and Jones (1998) investigated three factors that affect interaction with IR systems: functional visibility, cognitive load and balance of control between the searcher and system, relating them to a previous set of experiments. The functional visibility – allowing the searcher more information on how the system works – is important at two levels. Not only must the searcher be aware of what options are available at any stage but they must also be aware of the effect of these options. The study by Beaulieu and Jones demonstrated that interfaces such as the Checkbox system, that separate query modification and relevance assessment, can be more cognitively demanding for searchers. In this experiment subjects appeared willing to delegate responsibility for relevance assessment to the search system. However, they wished to retain control over query reformulation and retrieval strategy selection, activities they perceived as being important for the success of their search. That is, subjects were willing to delegate control over the provision of relevance information as long as they could control how this information was used.

A deeper understanding of what searchers want to control and what they are happy to delegate can assist in the development of more effective systems for interactive search. Techniques to facilitate the provision of relevance information, form new queries and use these queries were all tested in this experiment. The discussion of interface support is divided into three main parts: relevance indications, query creation and the selection of retrieval strategies. Each section begins with an italicised summary statement describing the main conclusion drawn.

12.3.1 Relevance Indications

Subjects wanted the search system to infer relevance. In all cases, systems that gathered relevance information unobtrusively from subject interaction were preferred to systems that required explicit subject involvement. Whilst the Checkbox system gave subjects an opportunity to directly indicate which items were relevant the additional responsibility dissuaded subjects from doing so. They felt that the implicit techniques were a reasonable approximation for their indications and were willing to delegate responsibility for this activity to the search system.

The Checkbox system differed from the other systems in how relevance information was conveyed; the subject was required to explicitly mark representations as being useful in their search. This was an onerous task that was not liked by subjects. In the experiment one subject commented "[checking boxes] added a new dimension to search that could become frustrating". This summarises the general opinion of experimental subjects; that the need to mark boxes was removed from the search for information and required a transition between two search activities. Subjects preferred systems that used implicit relevance assessments since they did not require them to mark items as relevant, they had difficulty marking items as relevant, they forgot to mark items and the marking of the items intruded in their searching. Implicit relevance assessments may not be as accurate as their explicit counterpart in determining which items are *definitely* relevant but they are able to build a larger body of evidence for those that are *potentially* relevant. The Checkbox system forced subjects to make binary assessments of what items were relevant; this may not always be appropriate as the relevance of a search result may be uncertain or partial (Spink *et al.*, 1998; Maglaughlin and Sonnenwald, 2002).

Experimental subjects tended to only mark items that were definitely relevant, meaning they did not provide the system with much evidence with which to make query modification decisions (i.e., only 2% of representations were marked). Techniques such as those employed by Aalbersberg (1992), Allan (1996) and Iwayama (2000) can be used to modify queries in situations where only a small amount of relevance information is available. 15 of the 48 experimental subjects suggested that the process of relevance feedback could also be improved if they could provide indications of what interface items or terms definitely were not relevant for their search. After they had given this negative relevance feedback they would not want to see items of this nature, or these terms, again during their search.

In this experiment 'precision' was taken as a measure of search effectiveness and based on how much of the retrieved document set the subjects classed as relevant. To compute this measure, the Checkbox system used the proportion of potential representations [45] that were actually marked and the implicit feedback systems used the proportion of all representations that were classified as being relevant. The results suggested a large difference between how much information the implicit systems regarded as relevant and what the subject actually marked as being relevant. The relevance and usefulness of the terms generated from the implicit feedback systems was higher than that of the Checkbox systems, suggesting that more evidence, albeit less reliable than that provided by the searcher allowed better quality terms to be chosen by the implicit feedback framework. It also suggests that criteria subjects employed when assessing relevance was too strict and that better queries could have arisen from the selection of more representations that were perhaps not totally relevant. In the next section I discuss the interface techniques used to incorporate new query words.

12.3.2 Query Generation

Subjects preferred to retain control over query creation. The systems that allowed subjects to monitor and change the query were preferred over the Automatic system, which did not. They were willing to delegate the task of recommending potential keywords but not the task of adding these words. Subjects preferred control over the terms chosen by the system, even if this meant more work for them in moving terms of interest from the recommended term list to the query. This effort was seen to be both *unnecessary* (subjects were not forced to do it) and *worthwhile* (subjects perceived a benefit from it). The implicit nature of the evidence captured may make the search decisions of systems that use it unreliable and subjects may rather retain control to be sure of their correctness. Subjects engendered more trust in systems where they could verify the correctness of the words chosen prior to their submission. For more complex tasks they required more support in query formulation.

Subjects liked having terms suggested to them, but in a way that did not require them to delete irrelevant terms (as in the Checkbox system), only select relevant ones; subjects did not want to have to act to correct erroneous system decisions. Subjects were more willing to delegate responsibility for the creation of queries to systems that allow them to verify the correctness of system decisions. In a related study, Koenemann and Belkin (1996) tested search systems with different levels of visibility and interactivity in creating queries. In this experiment the Automatic system only allowed subjects to see the query created by the system; the Checkbox and Recommendation systems allow subjects to view *and* adjust the new query. In this

[45] All document representations in the top 30 documents that could be marked.

experiment, as in the work by Koenemann and Belkin, subjects preferred systems that gave them control over the new queries. That is, they want help in selecting query terms but want ultimately to decide which terms are used.

The Checkbox system chose terms for subjects based on the items they had marked as relevant. These items reflected their current information needs and the terms suggested by the system appeared to reflect these needs also. Subjects chose terms from those recommended in the Recommendation system because: (i) they represented new ideas, (ii) they meant the same as the query terms, and (iii) they were related to the query terms. The study by Koenemann and Belkin found that subjects tended to choose semantically related feedback terms. In this experiment I found that subjects use the query terms to give them ideas for what terms are appropriate or were related to the original terms in some way. For example, a search for 'worldwide petrol prices' could mean that the terms 'pipe', 'iraq' and 'dollar' are good feedback terms, but their semantic relationship to the original query is not immediately apparent.

All experimental systems tried to increase the length of subjects' query statements by expanding the original search query. Belkin *et al.* (2003) have demonstrated that experimental subjects can be more satisfied with search results if they submit longer queries to the search system. The use of a feedback system to choose terms on a searcher's behalf is only one way to create longer queries. Kalgren and Franzen (1997) demonstrated that a different style of query input box encouraged the submission of longer queries, a result verified by the Belkin *et al.* (2003). It is preferable to encourage searchers to better define their information needs. However, in circumstances where they may be unfamiliar with the topic of the search, they may be unable to produce longer queries (Kelly and Cool, 2002).

Traditional Web search systems are 'pull' oriented where it is the searcher's responsibility to locate relevant information. The systems I have described in this thesis operate on a 'push' paradigm and are adaptive, work to better describe information needs and consider changes in these needs, restructuring or recreating the information presented at the results interface. Once a new query has been generated it can be used to perform a *retrieval strategy*. In the next section I discuss the selection of such strategies.

12.3.3 Retrieval Strategy Selection

Subjects preferred to retain control over search decisions. Systems that gave the subjects control over search decisions were preferred to those that did not. The Recommendation

system suggested decisions that subjects may execute. Subjects liked receiving this support but in a similar way to the creation of query statements wished to verify the correctness of any decisions before they were taken.

The Recommendation and Automatic systems dynamically update their internal representation of information need change and adopt the retrieval strategy to reflect the information need of the searcher, as estimated by the search system. Different search decisions had different levels of impact on a search. Reordering decisions restructured the already retrieved information at the interface, whereas re-searching decisions generated a new set of documents. The decisions increased in severity, from reordering Top-Ranking Sentences, to reordering documents, to re-searching the Web. Subjects appeared more willing to retain control over the number of re-search operations, but were willing to experiment with reordering. This suggests an association between the severity of the decision and subject's willingness to retain control over them. That is, for less severe strategies subjects were more willing to delegate responsibility to the system.

The implicit feedback frameworks evaluated in this thesis are dependent on how results are presented and how searchers interact with them. In the next section I discuss the presentation of information at the results interface and aspects of subject interaction.

12.3.4 Presentation and Interaction

In all experimental systems subjects suggested that they tried to look at information related to the search task. This was an important aspect of the experimental systems that used implicit feedback since they relied on subjects using the interface components as feedback on what information is relevant. It has been well documented that searchers will demonstrate a variety of information seeking behaviours during the course of a search (Ellis, 1989; Hancock-Beaulieu, 1990; Kuhlthau, 1991), and indeed will exhibit different kinds of interaction with different texts according to different goals, knowledge and intentions. However, searcher interaction is generally driven by a desire to maximise the amount of relevant information they view (maximise recall), whilst also minimising redundancy (maximise precision). Through monitoring the information they interact with I have shown that search systems can approximate subject's information needs.

The direct involvement of the searcher in the information seeking process results in a dialogue between them and the IR system that is potentially muddled and misdirected (Ingwersen, 1992). The systems described in the later parts of this thesis implement aspects of the

principle of *polyrepresentation* (Ingwersen, 1994) that suggests one should provide and use different cognitive structures during acts of communication to reduce the uncertainty associated with interactive IR. The cognitive structures around which polyrepresentation is based are manifestations of human cognition, reflection or ideas. In IR the author's text, including titles and the full-text are representations of cognitive structures intended to be communicated. However, these portions of text demonstrate different functional origins. That is, they have the same cognitive origin but were created in a different way or for a different purpose. Subjects generally responded well to the content-rich interfaces and suggested that the multiple document representations allowed them to focus on the most relevant parts of the documents. Some subjects remarked that they would like to be able to jump between steps in a relevance path. For example, in the search interfaces presented in Chapter Ten a searcher cannot move straight from a top-ranking sentence to that sentence in its source document context. This rigidity of the relevance path structure is a necessity of the implicit feedback model deployed (which is path based). The Binary Voting Model, described in Chapter Six, does not place such constraints on path traversal and would perhaps be more suited for search interfaces that wish to implement a less rigid term weighting methodology.

Overall, the findings suggest that subjects want to retain control over the strategic aspects of their interaction. That is, over the aspects that will directly influence the quality of the results offered or future directions of their search. They view the provision of relevance indications only as an operational activity required to receive assistance. There is a disparity between how important subjects regard the communication of relevance information and its importance to the search system. Although relevance feedback can be useful tool to improve search effectiveness, it is under utilised because of the interface techniques it uses to gather relevance information. To cater for this, search systems must incorporate new techniques for gathering relevance information. Implicit relevance feedback methods such as those described in this thesis may be useful to address this problem. Further research is required in the development of search tools that incorporate implicit feedback techniques for gathering relevance information.

In the next section results relating to the next research question – the effectiveness of the information need detection component – are discussed.

12.4 Information Need Detection

Searchers may have problems choosing terms to adequately represent their information needs (Taylor, 1968). In this thesis approaches for choosing terms to create new, improved queries are presented and evaluated with human subjects and a novel simulation-based evaluation methodology. In this section I discuss experimental findings on the information need detection part of the implicit feedback framework. This experiment tested the term selection component of the framework from the subjects' perspective in a series of information seeking scenarios on different experimental systems. The simulation-based study in Chapter Eight allowed me to benchmark the performance of the term selection models with simulated searchers. The success of the Jeffrey's Conditioning Model meant it was selected to choose terms for query modification in this experiment.

The same model was used in three interfaces and differences in subjects' perceptions of the relevance and usefulness of the terms were noticed between systems. This suggests that the way the terms are presented plays an important part in how the terms are perceived, independent of their value. Subjects were asked to assess the 'relevance' and 'usefulness' of the terms suggested by the framework. In task-oriented evaluations one would expect relevance to be synonymous with 'utility' (Cooper, 1973) or 'pertinence' (Saracevic, 1996), resulting in a strong correlation between relevance and usefulness. However in the evaluation there were statistical differences between the relevance and usefulness scores for five of the six system-group comparisons and overall among all subjects and all systems; subjects generally regarded terms as being more relevant than useful. This could be because subjects did not know what relevance was or they did not associate it with usefulness. Five of the 48 subjects commented on the difference between relevance and usefulness; they could recognise which terms are related to the search (topically relevant) but not which were useful in pushing the search forward in terms of changing search focus or retrieving more relevant documents (useful). So although they can recognise easily that terms are on topic they may have trouble saying which were useful. This example demonstrates the importance of asking the right questions in user experiments such as this. There is a danger that experimenters would typically ask whether the terms selected by the system are 'relevant' or 'useful', but not both. In doing so they would miss the distinction one can make between the two attributes.

Subjects assessed the usefulness of terms on a five point semantic differential, between 1 and 5 (inclusive). The lower the score assigned the more useful the terms. Overall, across all systems and subjects, the terms chosen by the system were assigned an average score of 2.18. This score was worse than one, the lowest (best) possible value. In Pilot Test 1 subjects did

not rate their own search terms as *always* useful, they acknowledge that they are not able to adequately conceptualise their information need, even when given the chance to refine the terms used to express it. However, as they view and process information, and their state of knowledge changes, they become more able to express these needs. The term selection model learns in a similar way, training itself with searcher interaction to better define what is relevant. It is difficult for any feedback model to choose useful terms, especially if subjects cannot even regard the terms they choose as useful. Unlike the discussion of interface support mechanisms in the previous section there were no differences in the usefulness of terms selected by the model for different types of search tasks.

Search systems that use implicit feedback techniques typically make decisions on behalf of searchers to assist them in their search. To operate effectively, such systems need to gain the trust of those that use them. In this experiment subjects were asked to indicate the extent to which they would trust the three experimental systems to choose terms on their behalf. The results again indicated a preference for the Recommendation system even though the same term selection model was used in all systems; both groups of subjects associated a higher level of trust with the Recommendation system. The true level of trust in the information need detection component is best measured independent of subject groups and independent of experimental systems. The average differential was 2.12, suggesting that subjects trusted the term selection component. The finding suggests that how the system communicates its decisions impacts on the level of trust subjects have in it.

During their searches subjects added new terms to their queries. These terms originated in ideas from a number of sources: (i) the terms recommended by the system, (ii) the retrieved documents and extracted information, (iii) a combination of these first two, (iv) the task being attempted, and (v) the subjects' tacit knowledge. The ideas derived from their search can result in a change in the direction of the search or the refinement of the current query statement with terms that better express information needs or better fit with the vocabulary of the collection. The terms suggested by all experimental systems appeared useful to initiate new ideas with around 20% of all new terms coming from ideas given by terms selected by the system. Ideas for terms also came from other sources, such as the task description, although it is conceivable that subjects will not always have search description as carefully constructed as a simulated work task situation.

The findings show that in systems that removed searcher control (i.e., the Automatic and Recommendation systems), subjects were more likely to use the terms proposed to initiate new ideas and search directions. The Checkbox system was dependent on subjects marking

results as relevant, and as a consequence, the terms suggested were from items the subjects already knew were relevant. In situations where searchers may benefit from a change in search direction it may be better to gather feedback implicitly as this can provide insight into their general, rather than exact, interests. Systems that remove searcher control over creating queries may be most appropriate for encouraging new and potentially useful search directions. This can be helpful if the searcher is struggling with their search. Although the findings discussed in the previous section suggest that searchers want to retain control over the additional terms used, it may not be in their interests to do so, especially if they lack the experience to devise well-formed queries.

The findings also show that the amount of interactivity in how additional words were chosen influences where the words were chosen from. When given less control, subjects were more likely to use the system's words or other sources such as the task, tacit knowledge or previous search experience. However, subjects did not use the documents or extracted information as inspiration for new words. Subjects depend on the Automatic system to reorder documents and Top-Ranking Sentences; subjects did not have any control over those activities in that system. From this, I conjecture that when subjects could not manipulate the space in which they searched, they were less likely to use that space to assist them in constructing new queries.

In the Recommendation system subjects were given a longer list of terms so they could be more selective about what terms were added. Subjects confirmed that the difference in the results was not related to the larger number of terms shown by the Recommendation system, but to the nature of the interface. Subjects were asked a simple 'yes'/'no' question as part of the informal discussion that followed the task on the Recommendation system. They were asked whether the larger number of terms in this system had an effect on their perceptions of the terms suggested; 42 of the 48 subjects responded 'no'; those that responded 'yes' found terms at a low-ranked position in the recommended list useful in their search. Subjects associated their preference for the Recommendation system with their perceptions of the query terms, showing that presentation factors can affect subject perceptions of such terms. In this experiment, the longer lists of suggested terms in the Recommendation system had only a minimal effect. The query length was restricted to a maximum of ten terms and the average initial query length across all systems, subjects and tasks was 2.86 terms.

In each of the experimental systems subjects were shown the terms the system had selected for them. In the Recommendation and Checkbox systems they were given the option to edit their query (i.e., add or remove terms). The results showed that in both systems subjects

typically accepted around 65% of the top six terms offered to them; demonstrating the effectiveness of the information need detection component. The Recommendation system showed 20 terms to the subject and allowed subject to move terms from anywhere in this list into the new query. In the analysis the list was divided into four quartiles, each containing 5 terms (i.e., the same number as in the Checkbox system). The scrollable window was sized so that the top six terms were shown at any time. The results show that more than three-quarters of terms (76.29%) came from the first 10 terms offered by the system; showing that the term weighting estimated which terms subjects were interested in. There were differences between subject groups in the rank position of terms chosen from the recommended term list. Experienced subjects were more likely to accept terms that appeared lower down the ranking (in the range 11-15). This may be because these subjects are interested in pushing the search forward through changing search focus or retrieving more relevant documents. Terms lower down the ranking may not be completely relevant and may foster the generation of new ideas.

In the studies described in Part II the experimental systems did not display the revised query, only the effect of the retrieval strategy that used the query (e.g., the reordered list of Top-Ranking Sentences). Subjects in those studies suggested that it would beneficial to see the terms used to allow them to make better decisions about the decisions made by the systems. In this experiment and in Pilot Test 1 subjects were shown the effect of the retrieval strategy chosen by the system and the revised query it created. That is, the query and its construction became a more prominent part of the search process.

In this section the results relating to the information need detection component of the system. The results showed that subjects found the terms selected by the framework relevant and useful in their search and that they would trust the framework to select terms for them. The terms chosen by the framework played a part in helping subjects create new query statements or make search decisions. In the next section findings related to the third research question, about the effectiveness of the information need tracking component, are discussed.

12.5 Information Need Tracking

The dynamic nature of information needs has been well documented (Bates, 1989; Harter, 1992; Bruce, 1994). As the need evolves, becoming more understood by the searcher, the searcher's actions and strategies may also evolve and a retrieval system should be able to adapt dynamically to this change. As well as refining query statements, the probabilistic framework also provides a mechanism through which it can support such evolving searches. The traditional view of information seeking assumes a searcher's need it static and

represented by a single query submitted at the start of the search session. However, it may well be dynamic and could change to reflect the information viewed by the searcher. As they view this information their knowledge changes and so does their problematic situation.

In situations where a searcher's need is ill-defined and liable to change, Bates (1989) among others (Ellis, 1989; Kuhlthau, 1993b) has argued that it may be beneficial to first explore the information space in a multidimensional way, allowing searchers to understand their information need more clearly. The classic model of the IR involves the retrieval of documents in response to a query devised and submitted by the searcher. RF is an example of an iterative process to improve a search system's representation of a static information need. That is, the need after a number of iterations is assumed to be the same as at the beginning of the search; the aim of relevance feedback is not to provide information that enables a change in the direction of the search. In situations where the information need is vague or uncertain, information that searchers encounter is more likely to give them new ideas and consequently new directions to follow (Belkin *et al.*, 1993). At each stage searchers do not just modify the search terms used in order to get a better match for a single query, rather the information need (as well as the search terms used) is continually shifting, to various degrees.

Berrypicking (Bates, 1989) is a technique where the information required to satisfy a query is the culmination of the knowledge gleaned from documents examined during the search session (Belkin, 2000). The interface techniques used in this experiment (especially in the Checkbox system) encourage an information seeking strategy similar to berrypicking. Rather than viewing the full-text of documents and refining their own queries, searchers visit a variety of document representations and receive support in their query refinement from experimental systems. The search interface presents many representations of the same document, biased towards the initial search request. The Recommendation and Automatic systems observe the information seeking behaviour of the searcher and use the evidence it gathers to better define information needs and cater for changes in these needs. The presentation strategies are manifestations of the berrypicking metaphor. The Checkbox system allows fragments of information to be directly stored by the subject and used for query refinement. The Recommendation and Automatic systems make inferences about all the information viewed and selects retrieval strategies to suit the estimated degree of change.

Through monitoring the information stored or viewed by searchers, the framework generates revised query statements. It is the differences between the system's estimation of the information need as it generates these statements and its formulation near the beginning of the search that it uses to estimate the extent to which the need has changed. The framework

chooses between three possible strategies aimed to support the user as they search; re-searching, reordering the document list, reordering the top-ranking sentence list and no-action at all. The strategies decrease in severity and reflect the estimated degree of change. Re-searching constructs a new information space and reordering restructures retrieved information depending on the level of change.

All subjects were instructed before the experiment that the different strategies provided varying degrees of interface support and had an increasingly dramatic effect on reshaping the information space. They were not told that the control related in any way to shifts, changes or developments in their information need as I felt this may bias their perceptions of the component. Searchers adapted well to the need tracking, and seemed comfortable with choosing between the different retrieval strategies.

The Recommendation and Automatic systems chose or recommended retrieval strategies. They were asked whether the retrieval strategy the system selected reflected any changes in their information need. There was a relationship between subject responses and the task categorisation used in this experiment. In the high complexity task there was scope for change whereas in the low complexity search task there was little.

The low complexity task was encouraged relevant or focused information seeking; the high complexity task encouraged explorative or browsing behaviour. Although the underlying topic is the same the additional restrictions placed on the low complexity search make the propensity to elicit changes in the information a subject is looking for also lower. [46] The findings of the experiment suggest that the information need tracking component was effective for high complexity tasks. The experimental systems selected more retrieval strategies for these types of task than for the low complexity tasks since in tasks of lower complexity subject's information needs remained more or less constant throughout their whole search. The more complex the search task, the more support subjects required in making decisions that had a strategic impact on their search. The information need tracking component appeared to not only track changes in the information needs, but the frequency of detected changes (and severity of chosen retrieval strategy) could be used to measure task complexity. For example, the selection of many retrieval strategies by the system may suggest that the search is variable and the search task is complex.

[46] The high complexity task is unclear about what information is being sought, how to obtain relevant information and how subjects will know when they have found relevant information. In contrast the low complexity task is generally clear about what information is required, how to find information and how to assess relevance.

Subjects were asked to rate how much they trusted the Recommendation and Automatic systems to select retrieval strategies for them. Although the subjects reacted positively to the retrieval strategy selected (i.e., the overall Likert scale response was significantly less than the middle value of the scale), they did not trust the information need tracking component as much as the information need detection component. This is perhaps because the potential implications of trusting the system to re-search information repositories or restructure the information displayed at the search interface are more severe than the selection of the some erroneous terms. Inexperienced subjects trusted the Automatic system less than the Recommendation system and since both used the same approach the difference may only be attributable to the presentation of the strategy. The Automatic system removed more control than the Recommendation system, selecting action and executing them without searcher consent. Inexperienced subjects commented that they did not feel in control of their search on the Automatic system. Experienced subjects felt similarly although some remarked that the removal of control was also a removal of burden and make the search simpler.

In a similar way to systems such as I^3R (Croft and Thompson, 1987) and FIRE (Brajnik *et al.*, 1996), the experimental systems created for the experiment in Part IV are always distinctly subordinate to the searcher. That is, the searcher always has the option to reverse system decisions. In Pilot Test 1 a search interface similar to that used in these experiments gave subjects the option to accept or reject search decisions after they occurred. In that experiment subjects commented that the communication of acceptance should be implicit as there was no need to tell the system they were happy with its decisions. In light of these comments a design decision was taken in the development of later experimental systems to only provide subjects with the option to 'undo'. Interaction logs were used to analyse the proportion of occasions that subjects reversed system decisions; around 70% of the search decisions made by the systems were accepted by subjects.

Some subjects commented that they would have liked to be shown a more comprehensive history of their search activity during their search including retrieval strategies chosen, queries submitted and all search results considered to be relevant. They also commented that they would like to be able to undo more than the previous action. In contrast, the experimental systems described in Part II did not provide any explicit notifications that search decisions had been made by the system or the option to reverse these decisions. In these systems a change in the rank order of the Top-Ranking Sentences was the first, and only, indication that the system had made a decision.

I measured the amount of overlap between the strategy chosen by the Checkbox and Recommendation systems and the strategies chosen by experimental subjects. A good information need tracking component should be able to predict searcher's decisions based on the variability of their search. This is a potentially difficult task and the reported success rate of 59.73% (almost 67% in the Recommendation system) appears reasonable. This was improved to 85.48% if I allowed some margin of error to include the search decision made and the next nearest decision (e.g., reorder Top-Ranking Sentences *and* reorder documents or reorder Top-Ranking Sentences *and* no action). The information need tracking component appears to make decisions that are appropriate for subjects' searches. In the next section I summarise the discussion presented in this chapter.

12.6 Chapter Summary

The results of the experiment show that it is possible to get searchers to interact with more than a few search results. The approach moves away from simply presenting titles to presenting alternative access methods for assessing and targeting potentially relevant information. From observations and informal post-search interviews across a series of related studies, subjects appeared to find the increased level of content shown at the results interface of value in their search. This is important, as the success of all experimental systems presented in this thesis – especially those that used implicit feedback techniques – is dependent on the use of these interface features.

The experiment tested different techniques for communicating relevance, creating queries and using these queries in different ways. Three experimental systems were developed that varied levels of control over each of these search aspects. These systems investigated which activities subjects wished to retain control over, and how much control they actually required. The results showed that searchers are happy to delegate full responsibility for indicating which search results are relevant, but only want to receive assistance in the formulation of query statements and selecting interactive search strategies. Subjects still wish to retain control over search activities they regard as important to the effectiveness of their search. Rather than trying to force searchers to provide feedback, implicit feedback techniques can remove the burden of indicating relevance, allowing subjects to focus on those activities they regard as important.

I found that the task categories used in the experiments were identifiable by subjects. That is, the variations in the task complexity were noticed by subjects even though they were not told that the complexity of the tasks differed. Subjects preferred the Recommendation system and

found it better for more complex tasks where more control over the query terms was preferable. The Checkbox system was good for the low complexity tasks where the objective of the search was clear. The Automatic system was good for complex searches where the subject did not want to be actively engaged in the information seeking process or may lack insufficient knowledge about the retrieval environment to choose the good terms. In general, the systems communicated with searchers in a way that was helpful.

The terms selected for query modification were both useful and relevant. Subjects did not correlate relevance with usefulness suggesting that they interpreted them as being two different things in their search. The approach tracked potential changes or developments in the information need based on changes in the document representations viewed by the searcher. The system communicated its prediction of these changes through the search decisions it made on the subjects' behalf. The retrieval strategies chosen by the system were appropriate and liked by subjects.

The success of the implicit feedback frameworks and the interface support mechanisms bodes well for the construction of effective search systems that use techniques to work in concert with the searcher. To approximate current needs the techniques presented do not use traditional, potentially unreliable (Kelly and Belkin, 2001), implicit sources of searcher preference (e.g., document reading time, scrolling), but interaction with granular document representations and paths that join them. Unobtrusively monitoring searcher interaction with content-rich interfaces such as those presented in this thesis may provide a means by which the potential of implicit feedback can be realised.

In Part V I present the conclusions drawn from the research presented in this thesis and avenues for future work.